

(SPSS Statistics 丛书)

SPSS 18 数据分析基础与实践

李洪成

编著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书主要介绍 SPSS 18 (中文版) 在数据处理中的应用, 结合实际案例来系统讲述数据处理和统计分析的方法与技巧。本书的统计学知识部分主要参照教育部《统计学》课程教学规范的要求。全书共分为十一章, 主要内容为 SPSS 统计分析软件简介、数据文件的建立、数据预处理、描述性统计分析、均值比较、非参数检验、相关分析、回归分析、方差分析、SPSS 输出管理和语法命令。本书各章基本上是各自独立的, 读者可以从第一章开始顺序阅读, 也可以选择感兴趣的章节进行阅读。

本书可以作为数据分析工作者的参考手册, 也可以作为高等院校《统计学》课程的实训教材, 或者作为 SPSS 统计软件的培训教材。

未经许可, 不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有, 侵权必究。

图书在版编目 (CIP) 数据

SPSS 18 数据分析基础与实践 / 李洪成编著. -- 北京 : 电子工业出版社, 2010.7
(SPSS Statistics 丛书)
ISBN 978-7-121-11255-3

I. ①S… II. ①李… III. ①统计分析—软件包, SPSS 18 IV. ①C819

中国版本图书馆 CIP 数据核字 (2010) 第 125580 号

责任编辑: 郭 立

特约编辑: 顾慧芳

印 刷: 北京天竺颖华印刷厂

装 订: 三河市鑫金马印装有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本: 787×1092 1/16 印张: 20.5 字数: 387 千字

印 次: 2010 年 7 月第 1 次印刷

印 数: 4000 册 定价: 45.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888。

质量投诉请发邮件至 zltz@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线: (010) 88258888。

前 言

2007年4月18日,国家统计局原定于该日上午10时召开第一季度国民经济运行情况发布会,但因为工作安排原因,发布会召开时间推迟至19日下午3时。传闻中的过热数据推迟公布,大大增强了市场对加息和收紧货币政策的预期。在4月18日的上海证券报上,有这样标题的文章:“一季度经济数据将公布,数据很悲观、后果很严重”。2007年4月19日,中国股市突然调整:沪指跌幅达4.52%,深指更下跌5.23%。很多分析认为,这一调整与国家统计局推迟公布宏观经济数据有很大的关系。4月20日,沪深股市强力反弹。今天,这样的情景经常出现,统计数据已经成为日常生活中大众所关注的重要内容之一。^①

随着各企业IT部门获得跟踪客户以及客户交易信息的许可和预算,经过几年的时间,许多大型数据仓库业已建立,现在许多企业积累的数据可能已经达到超大容量(PB或者EB以上级别)。金融、电信两个行业中拥有大量的客户数据和运营数据,它们很早就经营活动中大量应用统计分析的工具;制造业、零售业和政府等其他行业现在也开始大规模地应用统计分析工具。随着统计工具在经济生活中应用的深入,对统计方法和统计工具的要求变得更加迫切。

写作背景

SPSS 统计分析软件以其易用性和强大功能已成为目前最流行的统计分析工具之一。由于 SPSS 软件图形界面友好、上手快,故它是国内高校中讲授得最高的统计分析软件。尽管目前市场上介绍 SPSS 软件和相应统计方法的书籍很多,但其中大部分要么纯粹侧重于阐述统计方法,要么纯粹是 SPSS 软件文档的翻版,从中很难系统地学到 SPSS 软件的精髓和掌握它的应用。因此,笔者从教学工作的实际需要出发编著了本书。

^① 该段话引自 2009 年 8 月 22 日国家统计局教育中心主任张仲梁在第十二次全国中青年统计科学研讨会上的演讲:《独白的统计,对话的统计》。

本书内容

本书主要侧重于 SPSS 在实际工作中的应用, 结合实际的案例来系统讲述 SPSS 的相关应用。本书的统计学知识部分主要参照教育部《统计学》课程教学规范的内容要求。本书可以作为统计分析工作者的参考手册, 也可以作为《统计学》课程的实训教材, 或者作为 SPSS 统计软件的培训教材。全书共分为十一章, 第 1 章~9 章主要内容为 SPSS 统计分析软件简介、数据文件的建立、数据预处理、描述性统计分析、均值的比较、非参数检验、相关分析、回归分析、方差分析。实际工作中常常需要对分析的结果进行处理, 需要高效率运行统计分析, 因此第 10 章和第 11 章分别介绍了 SPSS 输出管理和语法命令。本书各章基本上是各自独立的, 读者可以从第 1 章开始顺序阅读, 也可以选择感兴趣的章节进行阅读。对于统计学的实训课程, 如果为两个学分, 可以选择如下: 第 4 章的描述性统计分析, 第 5 章和第 6 章的假设检验, 第 9 章的方差分析和第 7 章的相关分析。如果为三个学分, 可以选择第 4 章到第 9 章的全部内容。

本书特点

本书的每一部分都是从实际案例入手来讲述 SPSS 软件的实现, 对结果给出解释, 并对分析过程给出相应的建议。具体特点如下:

- 采用 SPSS 统计分析软件的最新版本—SPSS18 中文版, 介绍了最新的相关功能, 比如非参数检验、图表构建程序等;
- 从实际的案例入手, 大部分数据取材于 SPSS 自带的案例数据或者 SPSS 的培训案例;
- 对相应的选项给出详细的解释;
- 介绍了对实际统计分析工作者十分重要的 SPSS 输出管理和语法命令;
- 每章的后面给出了相关的思考和练习题, 读者可以对相关的内容和技巧进行练习。

感谢

在本书的写作过程中, 笔者始终得到了 SPSS 中国公司工程师们的支持, 他们是张宇客、徐晓琴、邵朱明、周嫦琦、李宁娜。其中, 张宇客提供了第 3 章和第 5 章的初稿, 周嫦琦提供了第 4 章的初稿, 李宁娜提供了第 8 章和第 9 章的部分资料。徐晓琴仔细阅读了本书的全部书稿, 提供了许多有意义的修改建议。姜政毅、曾凯、

曹文伟、李红军、赵鹏等也对本书的写作提供了帮助和支持。另外，本书的出版得到了电子工业出版社计算机分社社长郭立的大力支持，袁金敏和顾慧芳两位编辑付出了大量的劳动，在此一并衷心感谢！

技术支持网站

由于作者水平有限，书中难免有不足之处。如有偏颇之处，恳切希望读者批评指正。本书提供技术支持网站：books.minewin.com。在该网站，读者可以下载本书的案例数据、勘误信息。如果是采用该书作为教材的教师，可以下载本书的教学课件、习题答案、期中考试和期末考试的样题。

李洪成

2010年7月1日

目 录

第 1 章 SPSS 统计分析软件简介

1.1 SPSS 统计分析软件的发展	1
1.2 SPSS 版本和授权	2
1.3 SPSS 统计分析软件的特点	3
1.4 主要模块及功能简介	4
1.5 SPSS 的安装	8
1.6 SPSS 的几种基本运行方式	11
1.7 SPSS 的界面	13
1.8 SPSS 的图形用户界面	17
1.9 SPSS 帮助系统	19
1.10 小结	24
思考与练习	24
参考文献	25

第 2 章 数据文件的建立和管理

2.1 数据管理的特点	26
2.2 SPSS 数据编辑器简介	27
2.2.1 开始 SPSS	27
2.2.2 SPSS 的数据编辑器界面	28
2.3 新建数据文件、数据字典	33
2.4 保存文件	36
2.5 读入数据	38
2.5.1 读入 Excel 数据	38
2.5.2 读入文本数据	40
2.5.3 读入数据库数据	45
2.6 数据文件的合并	50
2.6.1 添加个案	51
2.6.2 添加变量	55

2.7 数据的拆分.....	60
附录：如何为数据库文件建立 ODBC 数据源.....	63
2.8 小结.....	65
思考与练习.....	66
参考文献.....	68

第 3 章 数据预处理

3.1 可视离散化.....	70
3.1.1 直接输入分割点.....	71
3.1.2 根据条件自动生成分割点.....	74
3.2 缺失值.....	78
3.3 数据校验.....	82
3.4 标识重复个案和异常个案.....	91
3.4.1 标识重复个案.....	91
3.4.2 标识异常个案.....	92
3.5 选择个案.....	100
3.6 小结.....	104
思考与练习.....	105
参考文献.....	106

第 4 章 描述性统计分析

4.1 频率分析.....	108
4.2 中心趋势的描述：均值、中位数、众数、5%截尾均值.....	111
4.2.1 均值.....	111
4.2.2 中位数.....	112
4.2.3 众数.....	112
4.2.4 5%截尾均值.....	113
4.3 离散趋势的描述：极差、方差、标准差、分位数和变异指标.....	113
4.3.1 极差（Range）.....	114
4.3.2 方差和标准差.....	114
4.3.3 变异系数.....	115
4.3.4 分位数.....	115
4.4 分布的形状——偏度和峰度.....	116
4.5 SPSS 描述性统计分析.....	117

4.5.1	频率入口	118
4.5.2	描述子菜单	120
4.5.3	探索子菜单	121
4.5.4	表格	123
4.6	应用统计图进行描述性统计分析	124
4.6.1	定性数据的图形描述	124
4.6.2	定量数据的图形描述	129
4.7	数据标准化	134
4.8	小结	136
	思考与练习	136
	参考文献	137

第 5 章 均值的比较

5.1	假设检验的思想及原理	138
5.2	均值	140
5.1.1	均值过程分析	141
5.1.2	双因素的均值过程分析	144
5.3	单样本 T 检验	145
5.3.1	数据准备	147
5.3.2	单样本 T 检验	148
5.3.3	置信区间和自抽样选项	150
5.4	独立样本 T 检验	151
5.4.1	数据初探	153
5.4.2	T 检验	156
5.4.3	均值差的绘图	159
5.5	配对样本 T 检验	160
5.6	小结	163
	思考与练习	164
	参考文献	165

第 6 章 非参数检验

6.1	非参数检验简介	166
6.2	单样本非参数检验	168
6.2.1	卡方检验	172

6.2.2	二项式检验	178
6.2.3	K-S 检验	188
6.2.4	Wilcoxon 符号秩检验	191
6.2.5	游程检验	192
6.3	独立样本非参数检验	193
6.3.1	独立样本检验简介	194
6.3.2	独立样本检验举例	196
6.4	相关样本非参数检验	198
6.4.1	相关样本检验简介	199
6.4.2	相关样本检验举例	201
6.5	小结	204
	思考与练习	205
	参考文献	205

第 7 章 相关分析

7.1	相关分析的基本概念	206
7.1.1	相关关系的种类	207
7.1.2	相关分析的作用	207
7.2	散点图	208
7.2.1	散点图简介	208
7.2.2	散点图——旧对话框	209
7.2.3	用图表构建程序绘制散点图	213
7.3	相关系数	215
7.3.1	线性相关的度量——尺度数据间的相关性的度量	215
7.3.2	Spearman 等级相关系数——定序变量之间的相关性的度量	221
7.3.3	Kendall 的 tau-b(K)	223
7.4	偏相关分析	223
7.5	小结	225
	思考与练习	225
	参考文献	226

第 8 章 回归分析

8.1	线性回归分析的基本概念	227
8.2	简单线性回归	229

8.2.1	简单回归方程的求解·····	230
8.2.2	回归方程拟合程度检验·····	231
8.2.3	用回归方程预测·····	233
8.2.4	简单线性回归举例·····	234
8.3	多元线性回归·····	236
8.3.1	多元线性回归方程简介·····	236
8.3.2	多元线性回归方程的显著性检验·····	237
8.3.3	应用举例·····	238
8.4	线性回归的诊断和线性回归过程中的其他选项·····	242
8.4.1	回归分析的前提条件·····	242
8.4.2	回归分析前提条件的检验·····	243
8.4.3	线性回归的其他选项·····	251
8.5	小结·····	254
	思考与练习·····	254
	参考文献·····	255

第 9 章 方差分析

9.1	方差分析的术语与前提·····	257
9.2	单因素的方差分析·····	257
9.2.1	描述性数据分析·····	258
9.2.2	单因素方差分析·····	259
9.3	多因素方差分析·····	264
9.3.1	多因素方差分析简介·····	264
9.3.2	多因素方差分析举例·····	265
9.4	协方差分析（ANCOVA）·····	270
9.4.1	协方差分析简介·····	270
9.4.2	协方差分析案例分析·····	271
9.5	小结·····	279
	思考与练习·····	280
	参考文献·····	280

第 10 章 SPSS 输出管理器简介

10.1	SPSS 结果浏览器简介·····	281
10.2	浏览和编辑 SPSS 的分析结果·····	282

10.2.1	目录区域的对象	282
10.2.2	内容区域	286
10.2.3	移动、复制和删除结果	287
10.2.4	添加和编辑文本	287
10.3	枢轴表编辑器	289
10.4	把表格转换为图形	295
10.5	打印输出结果	296
10.6	导出输出结果到其他程序	297
10.6.1	直接复制	297
10.6.2	导出为其他文件格式	298
10.7	小结	299

第 11 章 SPSS 编程简介

11.1	应用 SPSS 的五阶段	301
11.2	SPSS 语法简介	304
11.3	SPSS 语法编辑器	305
11.3.1	图形用户界面和语法编程相结合	305
11.3.2	SPSS 操作日志	306
11.3.3	SPSS 语法编辑器简介	307
11.4	应用 SPSS 语法命令进行编程	310
11.5	小结	311
	思考与练习	311

SPSS 统计分析软件简介

本章学习目标：

- 了解 SPSS 软件的应用情况和版本变化历史，明确 SPSS 统计分析软件的主要应用领域；
- 了解 SPSS 统计分析软件的特点；
- 明确 SPSS 主要模块及功能；
- 了解 SPSS 的安装方式和 SPSS 的界面，掌握 SPSS 运行的几种方式；
- 学习 SPSS 帮助系统。

1.1 SPSS 统计分析软件的发展

SPSS 是软件英文名称的首字母缩写，其最初为 Statistical Package for the Social Sciences 的缩写，即“社会科学统计软件包”。随着 SPSS 产品服务领域的扩大和服务深度的增加，SPSS 公司已于 2000 年正式将英文全称更改为 Statistical Products and Service Solutions，意为“统计产品与服务解决方案”，标志着 SPSS 的战略方向做出了重大调整。在 2009 年 3 月 19 日，SPSS 公司将 SPSS 四大系列产品（Statistics Family、Modeling Family、Data Collection Family、Deployment Family）整合到一个综合分析平台，把四类产品统一加上 PASW（为 Predictive Analysis Software 的首字母）前缀，喻义 SPSS 产品的发展方向为预测分析领域。此后 SPSS 把正在发行的 SPSS 17 统计分析软件正式更名为 PASW Statistics 17，此后发行的版本 18 的官方名称为 PASW Statistics 18。同年的 10 月 2 日，IBM 宣布完成收购 SPSS 公司，随后 SPSS 统计分析产品更名为 IBM SPSS Statistics。本书写作时用到的软件为 PASW Statistics 17，后来更新为 PASW Statistics18，也是目前的最新版本。虽然，产品名称历经变迁，但是软件自身的统计分析功能变化不大。在本书中，还是采用简单的名称：SPSS 或者 SPSS Statistics。

SPSS 是世界上最早的统计分析软件，是美国斯坦福大学的三位研究生于 20 世

纪 60 年代末开发出来的。Norman Nie——SPSS 软件的三位创始者之一，当时是斯坦福大学政治学的博士研究生，为了分析从多个国家收集到的几千份调查问卷，与 Bent（斯坦福大学运筹学方向研究生）、Hull 一起开发了一套自动化处理数据和输出统计分析结果的程序。他们开发的第一个版本于 1968 年正式发布。一开始，SPSS 就以其丰富的、高质量的文档而被广泛传播和应用。随着 SPSS 销售的迅速增长，SPSS 软件的两位创始人——Norman Nie 和 Hull 于 1975 年在芝加哥成立了 SPSS 公司。

1984 年 SPSS 公司首先推出了世界上第一个基于个人电脑的统计分析软件 SPSS/PC+，开创了 SPSS 微机系列产品的开发方向，极大地扩充了它的应用范围，并使其能很快地应用于自然科学、社会科学等各个领域。世界上许多有影响的报刊纷纷就 SPSS 的自动统计绘图、深入的数据分析、使用方便、功能齐全等方面给予了高度的评价与称赞。迄今，SPSS 软件已有 40 余年的成长历史。全球约有 25 万家产品用户，它们分布于通信、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业，是世界上应用最广泛的专业统计分析软件之一。

1.2 SPSS 版本和授权

截至目前（2010 年 3 月），SPSS 的最新版本为 PASW Statistics 18（尽管官方已经更名为 IBM SPSS Statistics，软件发行早于新名称，因此仍然沿用 PASW Statistics 名称）。SPSS 软件的升级相对比较有规律，基本上每年发行一个新版本。SPSS 软件的最近发行历史为：

- SPSS 11.0.1- 2001 年 11 月发布；
- SPSS 12.0 -2003 年发布；
- SPSS 13.0 – 2004 年发布；
- SPSS 14.0 -2005 年发布；
- SPSS 15.0.1 - 2006 年 11 月发布；
- SPSS 16.0.2 - 2008 年 4 月发布；
- SPSS Statistics 17.0.1 - 2008 年 12 月发布；
- PASW Statistics 17.0.2 – 2009 年 3 月发布；
- PASW Statistics 18.0 - 2009 年 8 月发布。

其中，SPSS 版本 14 和版本 16 都有中文版发行。从版本 17 开始，SPSS 把所有支持的语言集成到一起，可以在选项中选择十一种语言的任何一种版本。

SPSS 安装完成之后需要有授权号码才能正常运行。SPSS 17 有自带的试用授权，试用期为 1 个月。SPSS 18 的试用期限为三周。SPSS 程序安装完成后，会要求输入授权号码或者许可证。许可证授权向导允许您获取一个 PASW Statistics 许可证。如果您没有立刻获取许可证，可以启用临时试用许可证，临时许可证可以启用所有 SPSS 的高级模块。产品试用期从首次应用许可证开始，到期后，PASW Statistics 将不再运行，您必须获取许可证才能继续使用 PASW Statistics。许可证通过锁定代码（Lock Code）绑定计算机硬件。如果更换了计算机或其硬件，则需要新锁定代码并再次进行授权过程，才能继续使用 SPSS 软件。如果超出了许可证协议中规定的可允许授权数量，授权将失败。另外，许可证对时间变化敏感。如果更改了系统时间，将会导致 SPSS 软件运行失败。

注意：1. 必须获取许可证或启用临时试用许可证才能使用 PASW Statistics。
2. 安装正式授权软件之后，不要轻易更改系统时间，否则会导致软件运行失败。

1.3 SPSS 统计分析软件的特点

SPSS 是世界上最早采用图形菜单驱动界面的统计软件，它最突出的特点就是操作界面友好，输出结果美观。它将几乎所有的功能都以统一、规范的界面展现出来，使用 Windows 窗口展示出各种管理和分析数据的功能，以对话框方式展示出各种功能选择项。用户只要掌握一定的 Windows 操作技能，粗通统计分析原理，就可以使用该软件为特定的科研工作服务，或者进行企业级的数据分析。

SPSS 采用类似 EXCEL 表格的方式输入与管理数据，数据接口十分通用，能方便地从任何类型的数据文件或者数据库中读入数据。SPSS 统计过程既包括了常用的、成熟的统计过程，也包含了一些高级的统计分析方法，例如 Bootstrapping 方法、市场直销方法等，完全可以满足专业人士的工作需求。SPSS 输出结果十分美观，采用 SPSS 专有的 SPO 格式存储结果，结果也可以直接转存为 HTML 格式、文本格式、PDF 格式或者 PPT。对于熟悉老版本编程方式运行 SPSS 的用户，SPSS 还特别设计了语法生成窗口，用户只需在菜单中选好各个选项，然后单击【粘贴】按钮，就可以自动生成标准的 SPSS 程序。极大地方便了中、高级用户的使用。

SPSS 的主要特点如下。

（1）操作简单：除了数据录入及部分语法命令程序需要键盘键入外，大多数操作可通过“菜单”、“按钮”和“对话框”来完成。使用者只需要掌握简单的 Windows

操作技巧，便可应用 SPSS 软件进行统计分析。

(2) 无需编程：具有第四代语言的特点，只需告诉系统要做什么，无需说明要怎样做。只要了解统计分析的原理，而无需通晓各种统计算法，便可得到需要的统计分析结果。对于常见的统计方法，SPSS 的命令语句、子命令及选择项的选取大多可通过“对话框”操作完成。因此，用户无需花大量时间记忆大量的命令、过程、选择项，从而避免了漫长的学习过程。同时，熟悉或精通编程者，如果喜欢，可以通过编程来实现窗口和对话框分析的所有功能。

(3) 功能强大：具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。自带 11 种类型共计 136 个函数。SPSS 提供了从简单的统计描述到复杂的多元统计分析方法，比如：数据的探索性分析、统计描述、列联表分析、二维相关、秩相关、偏相关、方差分析、非参数检验、多元回归、生存分析、协方差分析、判别分析、因子分析、聚类分析、非线性回归、Logistic 回归等。随着版本的更新，SPSS 的功能不断完善，比如：在 SPSS 16 版本中增加了神经网络模块，在 SPSS 17 版本中增加了 EZ RFM 分析，在 SPSS 18 版本中新增了 Bootstrapping 分析。

(4) 方便的数据接口：能够读取及输出多种格式的文件。比如：由 dBASE、FoxBASE、FoxPRO 产生的*.dbf 文件，文本编辑器软件生成的 ASC II 数据文件，Excel 的*.xls 文件等均可转换成可供分析的 SPSS 数据文件。同样地，能够把 SPSS 的图形转换为 7 种不同格式的图形文件。SPSS 的输出结果可保存为*.txt、PDF、Word、Power Point 或者 html 格式的文件。

(5) 灵活的功能模块组合：SPSS for Windows 软件分为若干功能模块。用户可以根据自己的分析需要和计算机的实际配置情况灵活选择。

(6) 与其他程序的无缝结合：SPSS 新版本可以调用开源统计分析软件 R 或者开源高级程序语言 Python 的功能模块，实现诸如支持向量机（SVM）、关联分析、偏最小二乘等功能。

注意：SPSS 统计软件版本 17 和版本 18 更名为 PASW Statistics。IBM 购买了 SPSS 公司之后，统计分析产品的名称为 IBM SPSS Statistics。

1.4 主要模块及功能简介

SPSS Statistics 18.0 包含 17 个模块，这些模块的组合丰富了 SPSS Statistics 的

分析功能。这 17 个模块分别介绍如下。

（1）SPSS Statistics Core

SPSS Statistics Core 是 SPSS Statistics 软件运行的平台,确保您能综合使用 SPSS Statistics 软件的其他功能模块和 Statistics 家族的产品。在这个平台上,您可以完成任意需求的数据分析。该模块是从 SPSS 18 版本才开始有的,在以前版本中,它是和 SPSS Statistics Base 模块一起的。

（2）SPSS Statistics Base

SPSS Statistics Base 模块是 SPSS Statistics 软件的基础模块,它提供数据访问、数据管理和准备、数据分析和报告、统计图表等功能。含有基本的统计分析过程,例如计数、交叉列表分析、描述统计、探索分析、均值比较、方差分析、相关性分析、非参数检验、多重响应分析、因子分析、线性回归、曲线估计、聚类分析、判别分析以及尺度分析等。

（3）SPSS Regression

SPSS Regression 是非线性建模分析程序,使您能够应用高级、成熟的方法分析数据。当您预测行为和事件,而数据不满足线性回归假设时,可利用多项/二项 Logistic 回归、非线性回归、加权最小二乘、两阶段最小二乘、Probit 等回归方法实现。

（4）SPSS Advanced Statistics

SPSS Advanced Statistics 含有专门用以描述、拟合数据间内在复杂关系的统计方法,可以使分析更为准确,并得出更为可靠的结论。SPSS Advanced Statistics 提供了一组功能强大的单变量及多变量的高级分析技巧。SPSS Advanced Statistics 模块的多变量分析技术包括:广义线性模型(GZLMS)、混合模型、一般线性模型(GLM)、方差成分估计、MANOVA、Kaplan-Meire 估计、Cox 回归、层次对数线性模型、对数线性模型、生存分析等。

（5）SPSS Custom Tables

SPSS Custom Tables 在创建表格的同时,能够实时更新,帮助分析人员在较少时间里,做出美观、精确的表格。利用 SPSS Custom Tables 可以展现调查、客户满

意度、投票选举等的结果分析。灵活的交互功能，创建表格时的可预览性，及其统计推断和数据管理的能力，可以帮助用户方便清楚地了解结果。

（6）SPSS Categories

SPSS Categories 可以图形化展示数据中的潜在关系，通过启发性的概念映射、最优尺度、偏好尺度和数据降维技术，揭示数据中全部潜在的关系。SPSS Categories 为分析人员提供了深入分析复杂分类数据和高维数据的全部工具。通过 SPSS Categories 可以直观地解释数据，了解大型表中的计数、分级和排序中的关联情况。

（7）SPSS Exact Tests

为了确定变量之间的关系，研究人员往往首先查看交叉表或非参数检验中的 P 值。如果数据满足潜在的假设，用传统的计算方法是可行的。但是，如果数据属于小样本，或数据变量中很高的比例集中于某一类别，或者不得不将数据细分为几个类别，传统的检验方法将不能得出正确结论。SPSS Exact Tests 可以得到更为准确的结果。

（8）SPSS Missing Values

利用 SPSS Missing Values 填充缺失数据，建立更佳模型，得到更有效的结论。缺失值可能会严重影响分析结果，如果把它们忽略、或者计算时排除它们，很可能会得出不正确的结论。SPSS Missing Values 是每位关心数据有效性的分析人员的有利工具，该模块提供六种缺失值诊断报告，使分析人员可以从多个角度检查数据，发现数据缺失模式。然后，分析人员可以估计摘要统计量，并利用统计方法填充缺失值。

（9）SPSS Conjoint

SPSS Conjoint 是用来模拟消费者决策过程的研究工具。SPSS Conjoint 能加强对消费者偏好的理解，更有效地进行产品设计、定价和营销；帮助衡量产品或服务在消费者心中的位置。具备了这些知识，公司在设计产品时，就能把对于目标市场最重要的属性特征包含进来，根据这些属性值进行定价，并专注于最有可能吸引目标购买者的点上。即使市场上的竞争者、产品和价格随时间发生改变，依然可以继续利用由 SPSS Conjoint 得出的结果来模拟情况发生变动后的市场。这样在投入大量资源进行产品开发和营销活动前，就能够提前预测市场的响应。

（10）SPSS Complex Samples

SPSS Complex Samples 提供了专门的统计工具，帮助计算出复杂抽样设计的统计量和标准误差。绝大多数常规的统计软件都假定数据是通过简单随机抽样取得的，而简单随机抽样在大多数大规模调查中既不现实也不经济。此外，用常规统计分析方法分析此类样本数据有得到错误结果的风险。例如，统计数据的估计标准误差经常太小，易导致对精度产生错误的认识。SPSS Complex Samples 将抽样设计融合到调查分析中，因此能在由复杂抽样得到的总体中获得更加有效的统计推论。

（11）SPSS Decision Trees

SPSS Decision Trees 能识别群体及预测结果。SPSS Decision Trees 模块能够直接创建分类决策树，帮助快速并准确地识别群体，发现群体之间的关系并预测未来事件。可以应用分类决策树于分段、分层、预测、数据降维、变量筛选、类别合并，以及连续变量离散化等。高度形象化的图解以非常直观的方式展现分类结果，分析人员可以清楚地把分类结果给业务人员解释。这些树也方便探索结果，并直观地确定模型是如何展开的。直观的结果能够帮助分析人员找出具体的子群以及通过传统的统计方法难以发现的关系。

（12）SPSS Data Preparation

SPSS Data Preparation 可强化数据校验工作，从而获得更准确的分析结果。该模块使分析人员能够简单便捷地识别可疑或无效的观测、变量以及数据值；了解数据缺失的模式，总结变量的分布。SPSS Data Preparation 使数据校验效率化、流程化，简化了数据校验过程，可迅速地完成分析之前的数据准备，并使结果更为精确。

（13）SPSS Forecasting

该模块利用完备的时间序列模型提高预测能力，包括多重曲线拟合、平滑以及自回归方程估计。利用专家建模器，可自动从 ARIMA 和指数平滑模型中选择最佳拟合时间序列和因变量的模型，避免反复选择模型的工作。

（14）SPSS Statistics Adapter

企业用户通过 SPSS Adapter 可获得管理分析资产和分析过程的强大能力。SPSS Adapter 使得 SPSS 统计分析产品可与 SPSS PES（现在已经更名为 CDS）平台整合在一起。这种企业级水平的应用为数据和模型提供了集成化、保密性强、可审查的

资源管理器。

（15）SPSS Neural Networks

SPSS 神经网络模块可用来建模数据中复杂的输入输出之间的关系或者数据之间的模式。可以选择分类算法（分类输出）或者预测算法（数值型输出），目前可用的两类算法是多层感知器学习算法和径向基函数(RBF)。

（16）SPSS Direct Marketing

SPSS EZ RFM 基于最近购买（Recency）、频率（Frequency）、金额（Monetary）来细分消费群体，为市场营销者瞄准目标市场提供了所需工具。此类 RFM 分析以前是比较困难的，现在 SPSS 的市场直销（SPSS Direct Marketing）模块就可以方便地进行 EZ RFM 分析。

（17）SPSS Bootstrapping

SPSS Bootstrapping 模块可帮助创建更加可靠的模型，得到更加精确的结果。只要模型是稳定的，它就可以产生准确、可靠的结果。无论公共部门的学术、科研工作，还是企业的决策部门，bootstrapping 都是一种较好的检测模型稳定性的技术。SPSS Bootstrapping 模块使得这种技术变得非常简单，并且容易实现。SPSS Bootstrapping 通过重复抽样，快捷地估计出观测值的分布，估计标准误差和总体参数的置信区间，估计平均、中位数、比例、优势比、相关系数、回归系数以及其他统计量。并且 SPSS Bootstrapping 方法可以减小离群值和异常值的影响。因此，可以更加清楚地了解建模的数据。

另外，SPSS 从版本 17 开始提供了 R 编程支持。只要安装了 R 插件，在 SPSS 中就可以调用 R 的所有统计分析程序，这大大地扩展了 SPSS 统计分析软件的功能。

1.5 SPSS 的安装

从版本 17 开始，SPSS 软件不再是一种语言一个安装介质，而是所有的语言都集成到软件中，用户可以在各种语言之间自由切换。启动 SPSS 软件后，可以在英文、法语、德语、意大利语、俄语、日语、简体中文等十一种语言间切换。可以说，一套软件，可以选择任何主流语言。

SPSS 17 有两张 CD，而 SPSS 18 是一张 DVD；两者的安装步骤相同，主要步骤为：

(1) 插入安装盘, 打开光盘文件, 单击 `setup.exe`, 或者等出现安装界面时, 单击安装 PASW Statistics 18.0, 按步骤安装。

(2) 选择 license 类型安装。PASW Statistics 有三种许可证类型: 单用户许可、站点许可、网络许可。单用户许可证和站点许可证类似, 只是站点许可证可以允许 1 个以上的用户同时使用该许可, 而单用户仅允许一个用户使用该许可。如果超出允许的用户数目, 安装时会提示, 已经超出了许可证支持的用户数。网络许可证是安装 SPSS 软件到任意多台计算机, 而可以同时调用 SPSS 处理器的用户数目由网络许可证控制。这里我们选择单用户许可, 如图 1-1 所示。

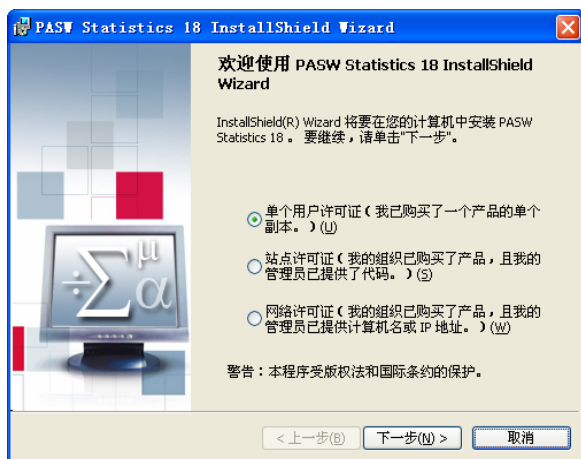


图 1-1 选择许可证

(3) 接受软件的许可协议

接受 SPSS 软件的许可协议条款的情况, 如图 1-2 所示。

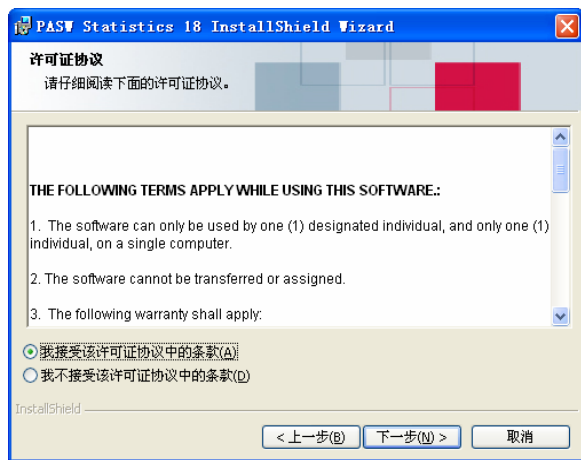


图 1-2 接受协议条款

(4) 选择安装的路径。默认安装路径为“C:\Program Files\SPSSInc\PASWStatistics18”。用户可以据情况修改安装目录，如图 1-3 所示。

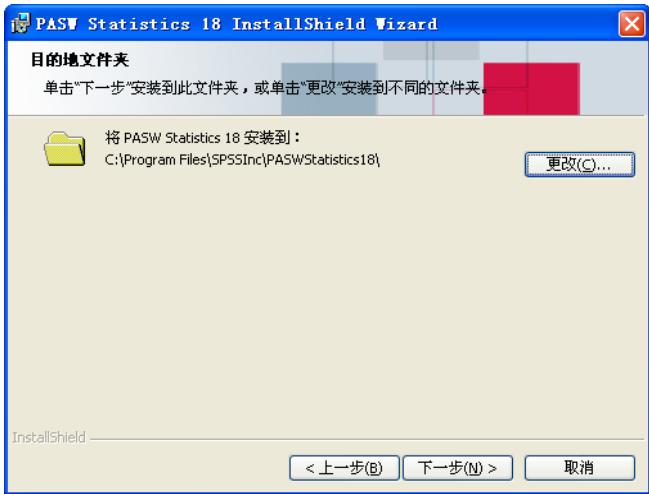


图 1-3 选择安装路径

(5) 单击“下一步”按钮进行安装。安装进程的情况如图 1-4 所示。

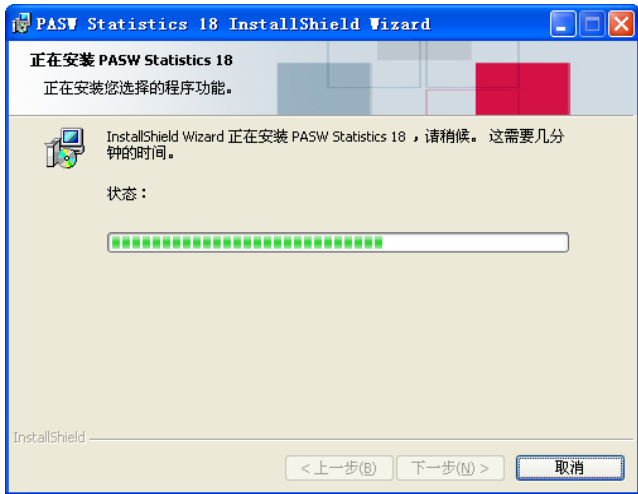


图 1-4 安装进程

(6) 选择产品授权方式。当安装进行到最后阶段，要求我们输入产品许可证。如果已经有了产品许可证，可以在下一步输入许可证号；否则，可以先启用临时试用许可证。SPSS 18 临时许可证允许 21 天的试用期，其选择产品授权方式如图 1-5 所示。

(7) 如果许可证输入正确，则成功完成安装。

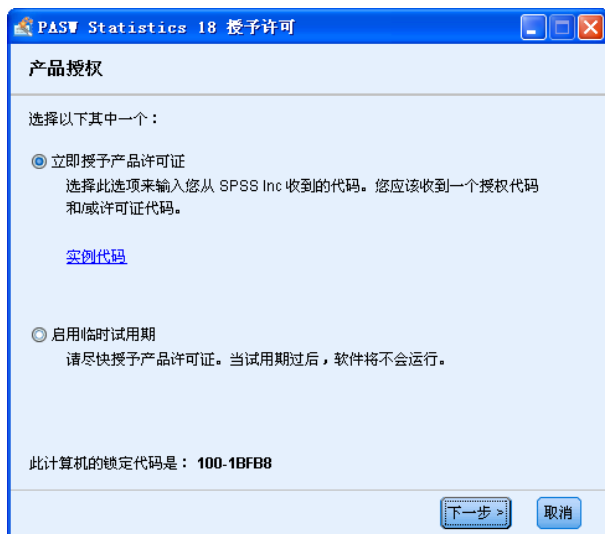


图 1-5 选择产品授权方式

1.6 SPSS 的几种基本运行方式

SPSS 提供四种运行方式。除了最常用的菜单操作方式，SPSS 还提供程序运行方式、Include 运行方式、Production Facility 方式。

(1) 菜单操作方式

SPSS 的常用操作方式是菜单操作。这种方法图形用户界面友好、操作简单、形象直观，能够一步步引导用户完成对数据的描述和模型的建立。例如，做一个频数分析。首先在菜单【文件】→【打开】中打开要分析的数据 Employee data.sav，对雇佣类别做频数分析。在“频率(F)”对话框中选择变量“雇佣类别”，如图 1-6 所示。单击【确定】按钮即可输出频数表，如表 1-1 所示。

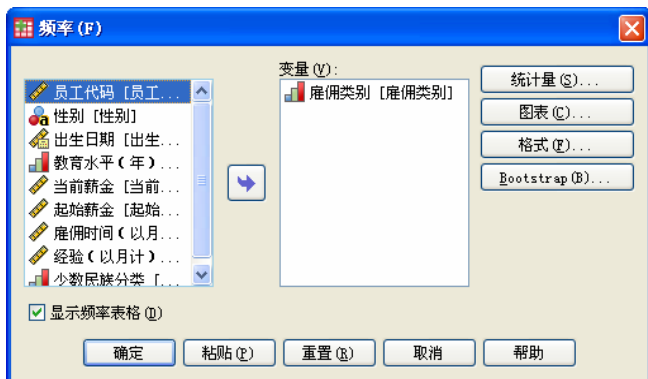


图 1-6 频率分析对话框

表 1-1 雇佣类别频率表

		频率	百分比	有效百分比	累积百分比
有效	职员	363	76.6	76.6	100.0
	保管员	27	5.7	5.7	
	经理	84	17.7	17.7	
	合计	474	100.0	100.0	

(2) 程序运行方式

SPSS 的程序运行方式是在 Syntax 编辑窗口中输入程序，如要打开一个数据 EmployeeData.sav，然后对雇佣类别做频数分析，则需要编辑如下程序：

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE='D:\SPSSIntro\Employee data.sav'.
DATASET NAME myData WINDOW=FRONT.
FREQUENCIES VARIABLES=雇佣类别
  /STATISTICS=VARIANCE RANGE MINIMUM MAXIMUM MEAN MEDIAN MODE
  SKEWNESS SESKEW KURTOSIS SEKURT
  /ORDER=ANALYSIS.
```

以上程序可以在任何文本编辑器中输入，也可在相应菜单操作的对话框中，用“Paste”按钮可以把相应的操作转化为 Syntax 语言。选择所有的语法命令行，单击“Run”工具按钮，就可以运行程序。或者在 SPSS 的语法编辑器窗口输入上述语法，如图 1-7 所示。

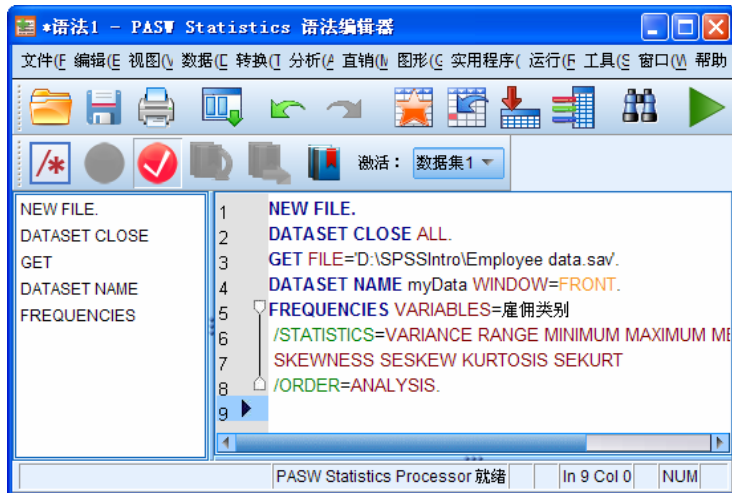


图 1-7 语法编辑器窗口

(3) Include 运行方式

在编写 Syntax 命令中，如果要调用其他语法文件时，除了复制粘贴现有的资源

外，还可以用 Include 的命令。如需要调用 automation.sps 文件，只要在 Syntax 编辑器窗口输入下列命令即可。

```
Include 'C:\SPSSIntro\Syntax\automation.sps'
```

（4）Production Facility 方式

Production Facility 生产作业方式提供了以自动化方式运行 SPSS Statistics 的功能。在生产作业方式下，程序可在无人看管的情况下运行，并在执行最后一条命令后终止，因此可以在其运行的同时执行其他任务或调度生产作业，使其在预定的时间自动运行。如果常常运行相同的一组耗时较长的分析（例如周报告，例行月度统计或者预测），则生产作业方式很有用。

生产作业使用命令语法文件告诉 SPSS Statistics 该做什么。在 SPSS 菜单中，选择【实用程序】→【生产工作】，得到如图 1-8 所示的生产工作窗口。



图 1-8 生产工作窗口

1.7 SPSS 的界面

SPSS 提供了五个窗口，分别为：数据编辑窗口、结果管理窗口、结果编辑窗口、语法编辑窗口、脚本窗口。

(1) 数据编辑窗口

SPSS 的数据编辑窗口是 SPSS 软件中常用的窗口，这个窗口主要用来处理数据和定义数据字典，它分为两个视图。一个是用于显示数据的数据视图（Data View）；另外一个为变量视图（Variable View）。

数据视图是用来显示数据集中的记录或者个案。它和 Excel 中的数据表十分类似。在数据视图中，一行代表一条记录（Case）或者一个个案，一列代表一个属性或者变量（Variable）。表头是变量名。在如图 1-9 所示的数据视图中，我们可以知道数据集的名字为“数据集 2”，它在物理上存储于数据文件“Employee data.sav”。另外，文件名的前面有一个星号（“*”），它表示当前的数据集刚刚做过修改，还没有保存。在数据视图中，我们可以修改数据，比如：修改已有的数据记录，删除记录，添加记录，或者修改一条记录的某一部分。其操作和 Excel 完全类似。

另外，从 SPSS17.0 开始，可以在数据视图对数据进行查找和替换。

员工代码	性别	出生日期	教育水平	雇佣类别	当前薪金	起始薪金	雇佣时间	经验	少数民族	变量	变量
1	m	02/03/1952	15	3	\$57,000	\$27,000	98	144	0		
2	m	05/23/1958	16	1	\$40,200	\$18,750	98	36	0		
3	f	07/26/1929	12	1	\$21,450	\$12,000	98	381	0		
4	f	04/15/1947	8	1	\$21,900	\$13,200	98	190	0		
5	m	02/09/1955	15	1	\$45,000	\$21,000	98	138	0		
6	m	08/22/1958	15	1	\$32,100	\$13,500	98	67	0		
7	m	04/26/1956	15	1	\$36,000	\$18,750	98	114	0		
8	f	05/06/1966	12	1	\$21,900	\$9,750	98	0	0		
9	f	01/23/1946	15	1	\$27,900	\$12,750	98	115	0		
10	f	02/13/1946	12	1	\$24,000	\$13,500	98	244	0		
11	f	02/07/1950	16	1	\$30,300	\$16,500	98	143	0		
12	m	01/11/1966	8	1	\$28,350	\$12,000	98	26	1		
13	m	07/17/1980	15	1	\$27,750	\$14,250	98	34	1		
14	f	02/26/1949	15	1	\$35,100	\$16,800	98	137	1		
15	m	08/29/1962	12	1	\$27,300	\$13,500	97	66	0		
16	m	11/17/1964	12	1	\$40,800	\$15,000	97	24	0		
17	m	07/18/1962	15	1	\$46,000	\$14,250	97	48	0		
18	m	03/20/1956	16	3	\$103,750	\$27,510	97	70	0		
19	m	08/19/1962	12	1	\$42,300	\$14,250	97	103	0		
20	f	01/23/1940	12	1	\$26,250	\$11,550	97	48	0		
21	f	02/19/1963	16	1	\$38,850	\$15,000	97	17	0		

图 1-9 数据视图

变量视图的功能是定义数据集的数据字典，它用来定义和显示数据集中的变量信息，变量视图如图 1-10 所示。

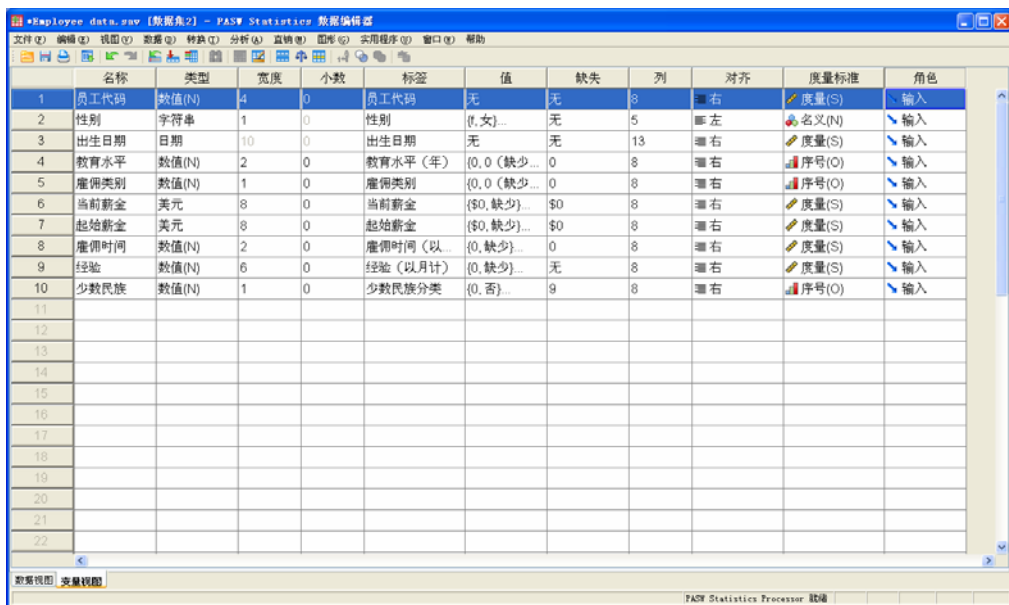


图 1-10 变量视图

(2) 结果管理窗口

SPSS 的结果窗口也称为结果视图或者结果浏览器，该窗口用于存放 SPSS 软件的分析结果，如图 1-11 所示。整个窗口分为两个区：左边为目录区，是 SPSS 分析结果的目录；右边是内容区，显示与目录对应的内容。

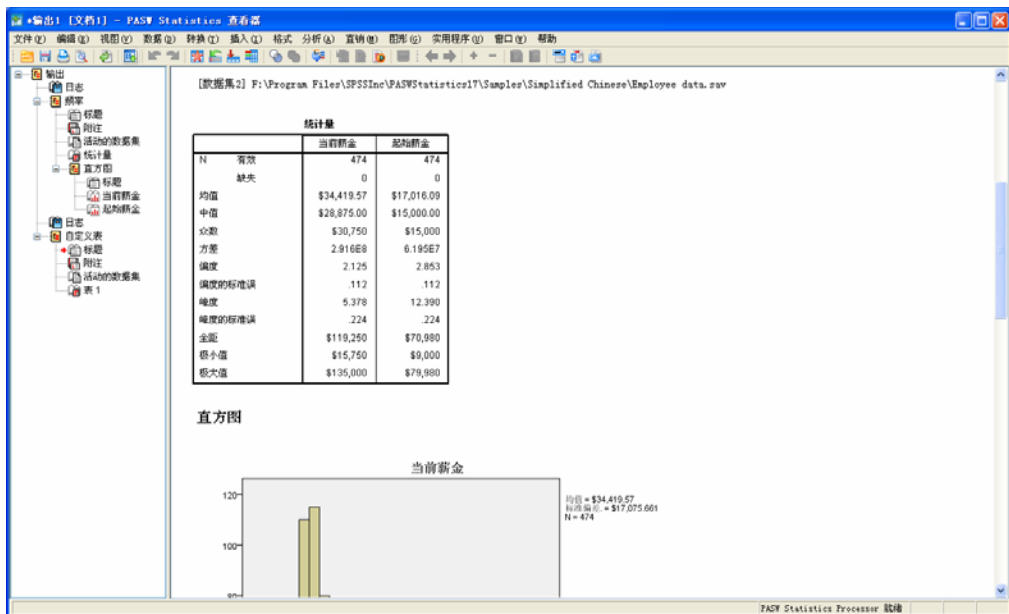


图 1-11 结果管理窗口

（3）结果编辑窗口

结果编辑窗口是编辑分析结果的窗口。在结果视图中，选择要编辑的内容，双击或者单击右键选择“编辑内容”，选中的图形会出现在“图表编辑器”中，这样可以在一个独立的窗口中编辑该图表。图表编辑器窗口如图 1-12 所示。

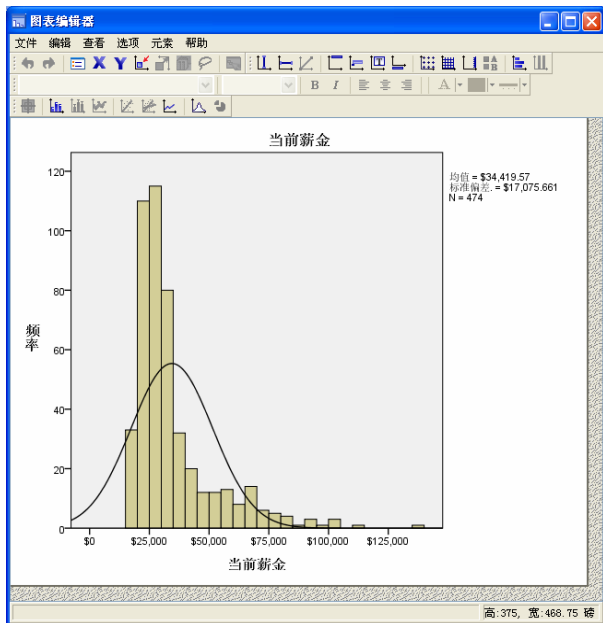


图 1-12 图表编辑窗口

（4）语法编辑窗口

SPSS 除了提供菜单操作，也提供语法编程方式。语法编程除了能够完成窗口操作所能完成的所有任务外，它还能够完成许多窗口操作所不能完成的其他工作，具体介绍参见本书第 11 章的“SPSS 编程简介”。另外，在语法编辑窗口，可以调用开源软件 R 中的任何程序。语法编程方式是对菜单功能的一个补充，它可以使繁琐的统计工作得以简化，特别是一些需要重复进行的工作。SPSS 编程是经常进行统计分析的人员和高级用户喜爱使用的方式，语法编辑窗口如图 1-13 所示。

（5）脚本窗口

SPSS 脚本是用 Sax Basic 语言编写的程序。脚本可以使 SPSS 内部操作自动化，可以自定义结果格式，可以连接 VB 和 VBA 应用程序，脚本编辑窗口如图 1-14 所示。

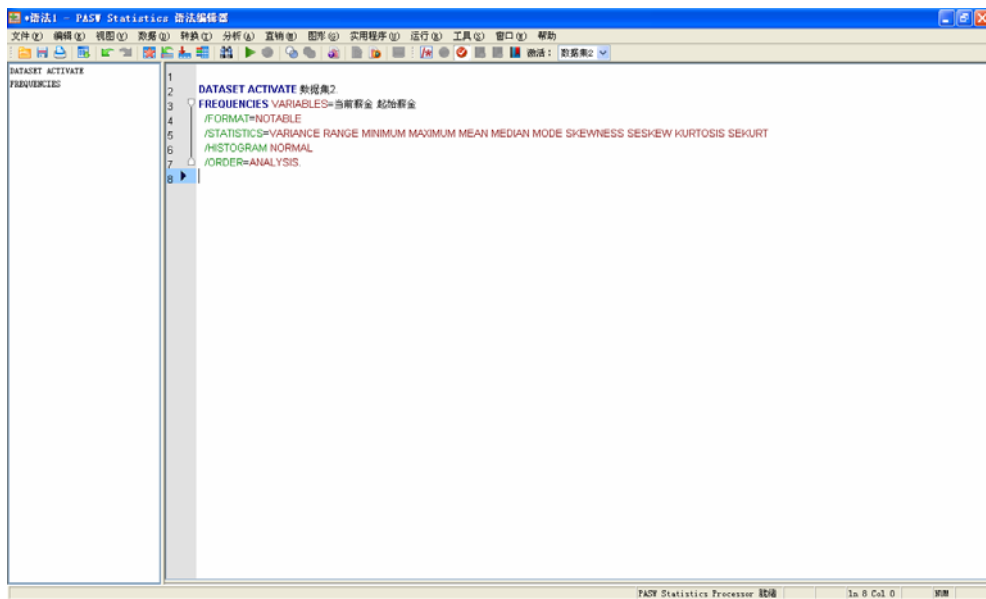


图 1-13 语法编辑窗口

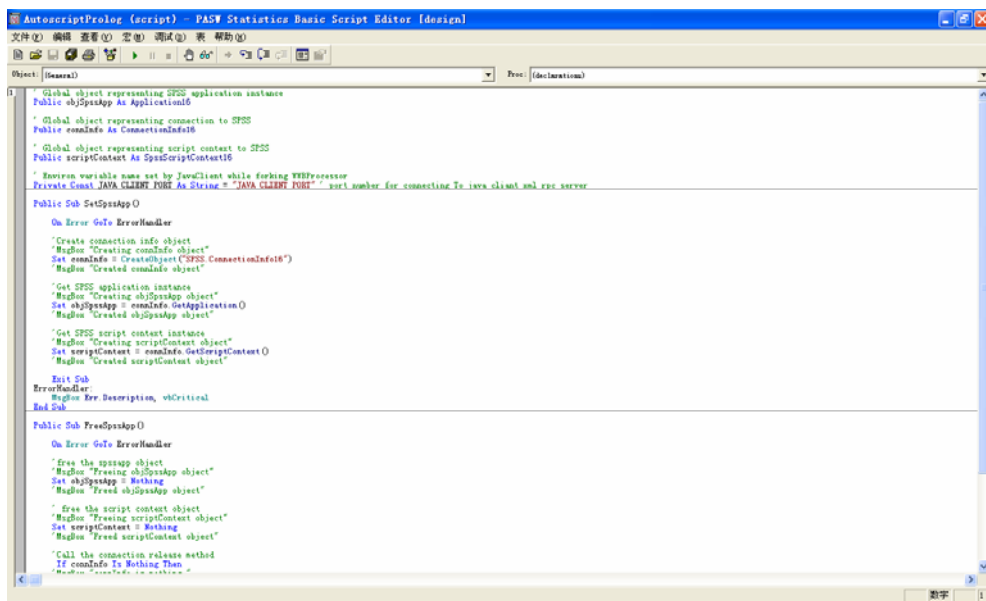


图 1-14 脚本编辑窗口

1.8 SPSS 的图形用户界面

SPSS 的统计分析功能主要通过三个图形用户界面来调用：他们分别为数据视图窗口、变量视图窗口和结果管理窗口。下面以数据视图窗口为例来简单介绍 SPSS 的菜单。SPSS 的数据视图窗口如图 1-15 所示，有 11 个菜单栏。其中，前三个菜单栏和 Excel 或者 Word 的前三个菜单名称一样，但是子菜单的内容有所区别。

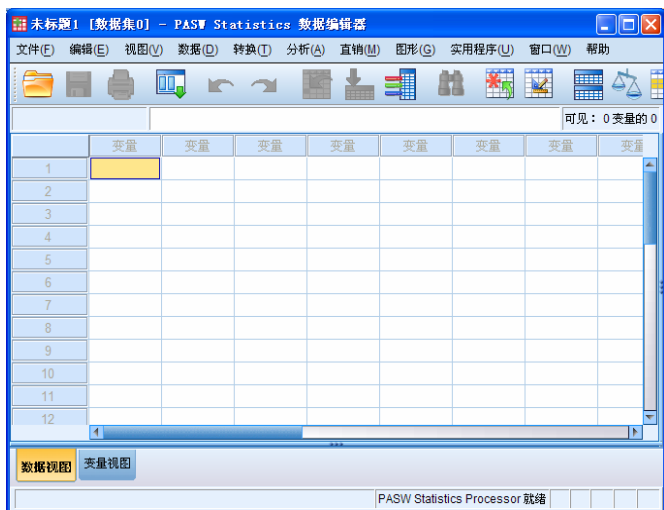


图 1-15 数据视图窗口

图 1-16 到图 1-20 分别是 SPSS 的数据菜单、转换菜单、分析菜单、图形菜单和实用程序菜单。

数据菜单主要完成数据字典的定义、排序、数据验证、合并文件等，不会改变原始数据。

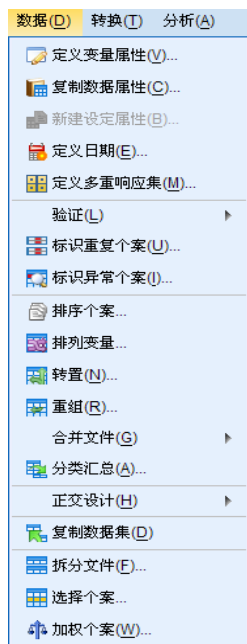


图 1-16 数据菜单

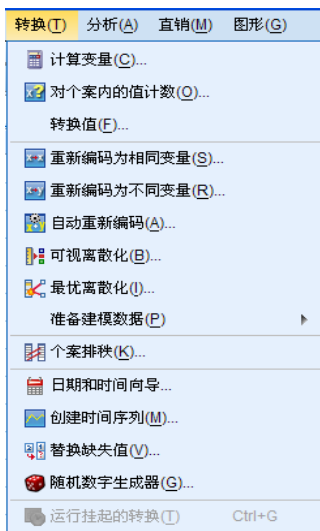


图 1-17 转换菜单

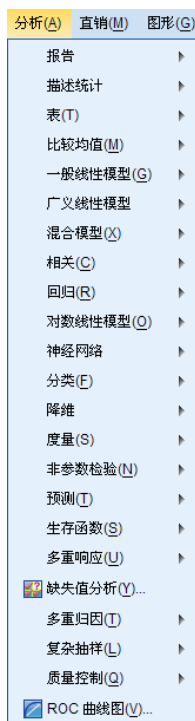


图 1-18 分析菜单

转换菜单的命令，主要用于基于原始变量重新转换生成新变量，如计算变量、

可视离散化、创建时间序列等，其中，计算变量是在数据预处理中应用最广泛的。

分析菜单包含了 SPSS Statistics 主要的统计分析功能，如所有的描述性统计分析命令、推断性统计分析命令，无论是一元统计还是多元统计分析，都集中在该菜单中。

图形菜单，如图 1-19 所示，用于生成各种统计图形，例如条形图、散点图、线图、面积图、直方图、箱图、饼图等。

实用程序菜单，如图 1-20 所示，主要提供了一些高效率的应用 SPSS 的方法，例如定义变量集，生产工作，集成 R 或者 Python 的外部程序等。

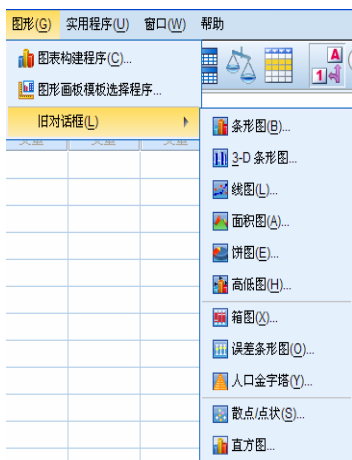


图 1-19 图形菜单



图 1-20 实用程序菜单

1.9 SPSS 帮助系统

SPSS 提供了友好的“帮助”功能，可以随时随地地为不同层次的用户提供帮助。其帮助系统包括主题帮助、教程、个案研究、统计辅导、语法命令参考、算法参考，以及 R 或者 Python 编程扩展帮助。另外，SPSS 系统的每个对话框都提供联机帮助。

(1) 主题帮助

SPSS 的主题帮助，在菜单上选择【帮助】→【主题 (P)】进入。SPSS 的主题帮助提供目录和索引两种方式查找所需的内容，如图 1-21 所示。

树形目录就像一本电子书，将所有主题组建成一个树状结构。只要单击左边的主题，就可以找到所需的内容。在索引方式中，只要在索引栏中输入关键词，系统就会展现相关的主题。



图 1-21 SPSS 联机帮助——主题

(2) 教程

SPSS 教程是为初级学者提供的学习资料，一步步指导学习者完成某些分析，以图形化、实例化的方式指导初学者如何使用 SPSS。初学者可以通过这个教程掌握 SPSS 的基本操作，如图 1-22 所示。

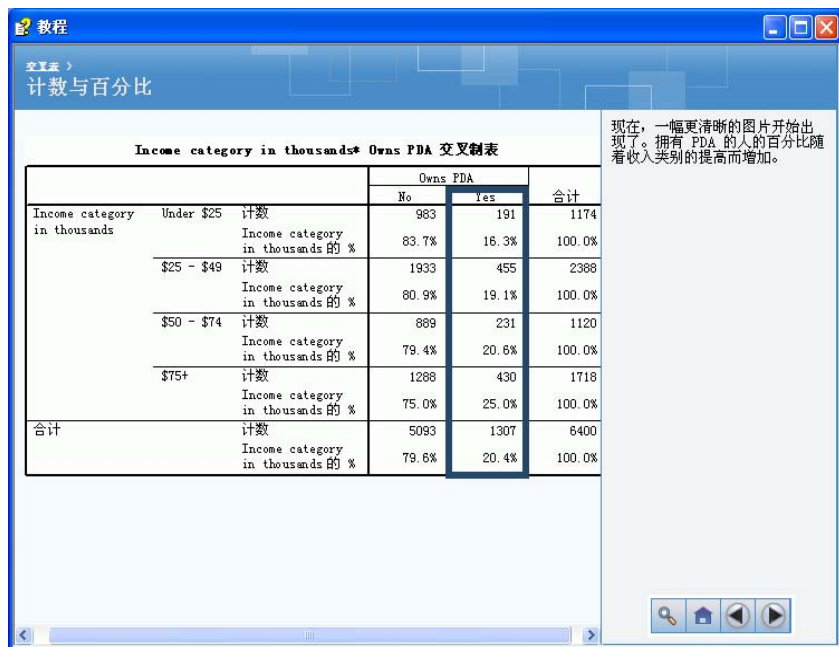


图 1-22 SPSS 帮助——教程

（3）个案研究

SPSS 的案例研究可以给中高级用户提供 SPSS 各模块主要分析方法的基本操作和结果的解读，其指导方式也是按图形化、实例化的讲解方式。虽然个案研究的语言目前是英文，但是通过这个指导教材，用户可以掌握绝大多数的 SPSS 操作，熟悉 SPSS 的各种高级分析功能及其应用背景，如图 1-23 所示。

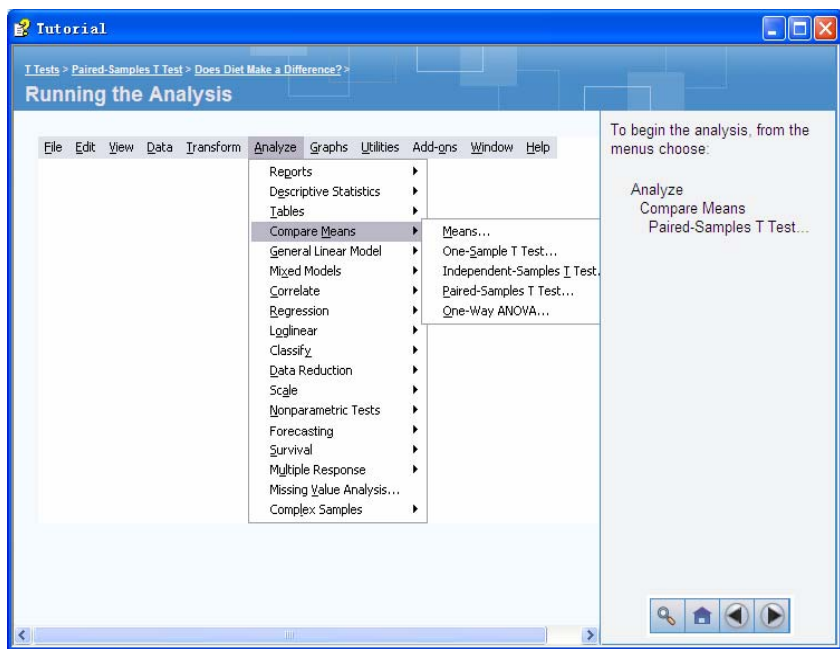


图 1-23 SPSS 帮助——个案研究

（4）统计辅导

SPSS 提供了统计辅导功能，它可以告诉用户为达到分析目的应选择什么统计分析，并且一步步地指导用户如何进行统计分析，如图 1-24 所示。

（5）语法命令参考

对于 SPSS 的高级用户来说，有时会发现对话框操作比较繁琐，甚至觉得某些高级功能用对话框无法完成。实际上，大约有 20% 的高级功能是必须要使用编程方式实现的，而且编程方式可以提高操作效率，处理更复杂的分析。为了方便高级用户的操作，SPSS 帮助系统提供了语法指南及相关分析的算法。只要选择【帮助】→【语法命令参考】，就可以打开相应的语法指令 PDF 文档。通过该参考文档，用户可以找到各种语法命令的用法，方便用户对高级操作的学习，如图 1-25 所示。

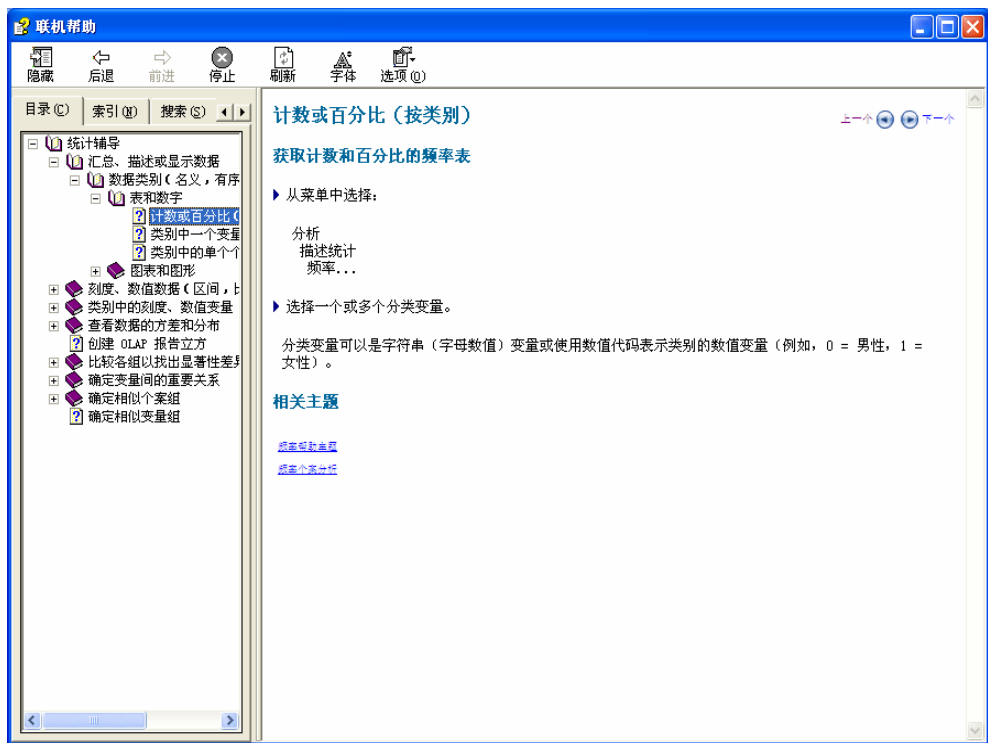


图 1-24 SPSS 帮助——统计辅导

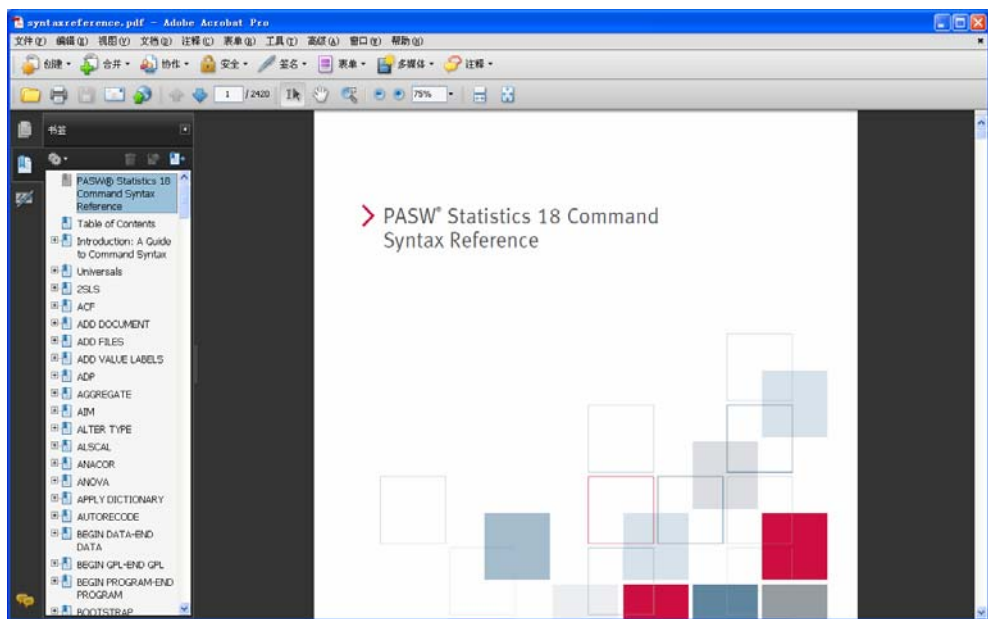


图 1-25 SPSS 语法命令参考文档

同时，SPSS 也对每个语法命令提供了联机帮助，在语法命令编辑器中，在相应的语法命令中，按 F1 键即可提供该语法命令的联机帮助，如图 1-26 所示。

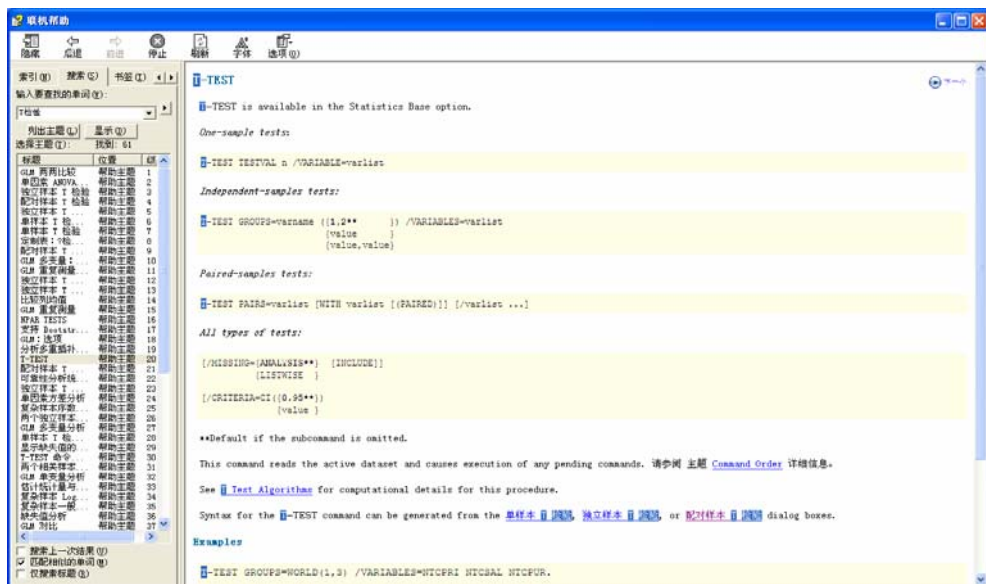


图 1-26 SPSS 帮助—语法命令参考

(6) 算法参考

选择【帮助】→【算法】，将出现 SPSS 算法的联机文档，如图 1-27 所示。该联机文档提供了 SPSS 的各种统计分析所采用的算法。通过该文档，高级用户可以了解 SPSS 统计过程应用的具体算法，加深对统计过程的理解。

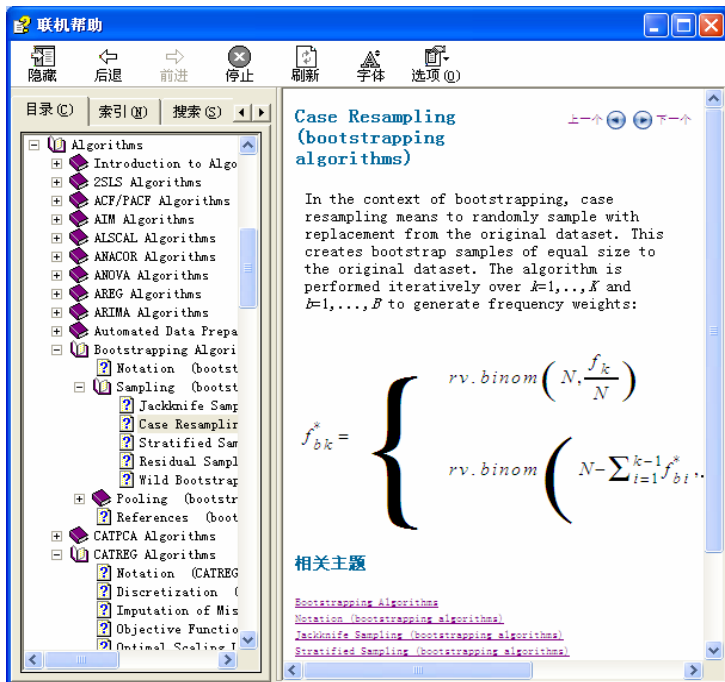


图 1-27 SPSS 帮助——算法参考

1.10 小结

本章主要介绍了 SPSS 统计分析软件的历史和特点。用户需要了解 SPSS 统计分析软件的模块化组成方式，同时了解运行 SPSS 统计分析软件的四种基本方式。应用 SPSS 进行统计分析需要用到五个窗口，读者需要熟悉这五个窗口的特点及作用。同时，用户还需要了解 SPSS 各个菜单大致所完成的功能，学会应用 SPSS 的帮助系统。

思考与练习

1. SPSS 发行版本的说法，正确的是：
 - A) 两年发行一个新版本
 - B) 一年发行一个新版本
 - C) 没有任何规律
2. 哪些是 SPSS 统计分析软件的窗口：
 - A) 结果查看器窗口
 - B) 枢轴表窗口
 - C) 决策树视图窗口
3. SPSS 帮助系统可以提供：
 - A) 算法指导
 - B) 语法命令参考
 - C) 根据统计分析主题组织的帮助系统
4. 下列哪些模块是 SPSS 18.0 的新增模块：
 - A) 回归分析模块
 - B) 自抽样模块
 - C) 神经网络模块
 - D) 市场直销模块
5. 哪些方式不是 SPSS 提供的运行方式：
 - A) 完全窗口菜单方式
 - B) 程序运行方式
 - C) 生产作业方式
 - D) 互联网运行方式
6. 哪些功能是 SPSS 基本模块 (Base) 所不能够直接实现的功能：

- A) 数据管理与准备
- B) 数据访问
- C) 统计分析
- D) 数据计划
- E) 数据收集

7. 哪些类型的文件是 SPSS 不能够直接打开的:

- A) *.sav 数据文件
- B) *.sys 数据文件
- C) *.dbf 数据文件
- D) SAS 统计软件产生的数据文件
- E) *.html 文件

参考文献

1. 维基百科全书: <http://en.wikipedia.org/wiki/SPSS>, http://en.wikipedia.org/wiki/Norman_Nie.
2. SPSS 公司网站: <http://www.spss.com>.
3. Norusis, M., SPSS 16.0 Advanced Statistical Procedures Companion. Upper Saddle-River, N.J.: Prentice Hall, Inc.,2008.

数据文件的建立和管理

本章学习目标：

- 了解 SPSS 数据编辑器特点，熟悉 SPSS 的变量视图和数据视图，掌握 SPSS 常用的工具按钮；
- 掌握数据录入 SPSS 软件的方法；
- 掌握把电子表格、数据库、文本文件等格式的数据文件读入 SPSS 软件的方法；
- 掌握 SPSS 数据集的数据字典；
- 明确分割 SPSS 数据文件的方法；
- 学习合并两个数据文件的方法。

“数据”是数据分析、统计检验与预测分析的基本对象，就像厨师手中的蔬菜、调味品等入锅的材料，它是统计部门、市场调查等部门所面对的对象。数据采集完成之后，需要录入软件分析系统才能够进行分析和应用。本章主要学习 SPSS 录入数据的特点，SPSS 的数据字典，如何把原始数据录入 SPSS 软件，如何把其他系统生成的数据文件导入到 SPSS 软件中。另外，本章还介绍如何把多个数据文件中的数据合并为一个数据文件，如何根据组别把数据文件拆分成若干组。

2.1 数据管理的特点

SPSS 数据管理器与 Excel 电子表格十分相似，所见即所得。SPSS 数据编辑器的每一行数据称为一个个案（Case），对应一个对象的记录，该对象可以是一个员工，一个样品，等等。每一列数据代表个体的属性，即变量（Variable）。例如，一个职工有员工号，姓名，职务，部门，联系电话，工资等属性，它们分别对应于 SPSS 数据集中的一个变量。在 SPSS 数据编辑器中，该职工的员工号、姓名、职务、部门、联系电话、工资等分别在相应的列上。用户可以直接在数据视图对数据进行修改。例如，可以直接在数据视图进行复制、粘贴，也可以直接修改某个个案的

属性值，或者删除某一行或者某一列，添加一列（插入变量）或者一行（插入个案）；另外，还可以进行查找、（批量）查找替换等。

SPSS 的主要文件类型如下。

SPSS 数据文件的默认格式为*.sav。SPSS 16 以及以后版本的结果文件默认格式为 *.spv。SPSS 统计分析的结果可以用文件的形式保存下来，SPSS 版本 15 或者以前版本的结果文件格式为*.spo，在其以后的版本中不能直接打开，需要安装结果浏览器软件—Legacy Viewer。

2.2 SPSS 数据编辑器简介

2.2.1 开始 SPSS

当启动 SPSS 软件（SPSS Statistics）以后，默认情况下首先弹出如图 2-1 所示的 SPSS 开始界面对话框。如果选择左边部分的两个选项，你可以进行如下的选择：

- 选择【打开现有的数据源】，打开最近使用过的数据文件；
- 选择【打开其他文件类型】，打开最近使用过的其他类型的非 SPSS .sav 格式的文件，例如 SPSS 语法文件 (*.sps)，SPSS 结果输出文件 (*.spv) 等。



图 2-1 SPSS 开始界面

或者选择右边的 4 个选项之一：

- 选择【运行教程】，将出现 SPSS 统计分析软件的教程，你可以从中系统地学习 SPSS 统计分析软件的各项功能；

- 选择【输入数据】，可以输入全新的数据；
- 选择【运行现有的查询】，可以运行已有的 Sql 查询语句，在 SPSS 数据编辑器中显示查询结果；
- 选择【使用数据库向导创建新查询】，SPSS 数据库向导将帮助你一步一步地从数据库中获取数据。

如果勾选图 2-1 下方的“以后不再显示此对话框 (D)”，则以后启动 SPSS 软件时，图 2-1 所示的对话框将不再出现。

2.2.2 SPSS 的数据编辑器界面

1. 数据编辑器界面

SPSS 数据编辑器有两个界面，数据视图界面和变量视图界面。数据视图界面的数据编辑区是数据的信息，而变量视图的数据编辑区是变量的信息。变量视图界面除不含编辑区选择栏外，其他和数据视图类似。

SPSS 的数据视图和 Excel 电子表格相似，其操作也类似 Excel，绝大部分数据管理和分析工作可以通过图形用户界面来完成，例如，你可以直接输入数据；可以单击鼠标右键，对数据进行“复制”，“粘贴”，“剪切”，“拼写检查”等。在图 2-1 中，我们选择【输入数据】，将得到如图 2-2 所示的数据编辑器窗口。

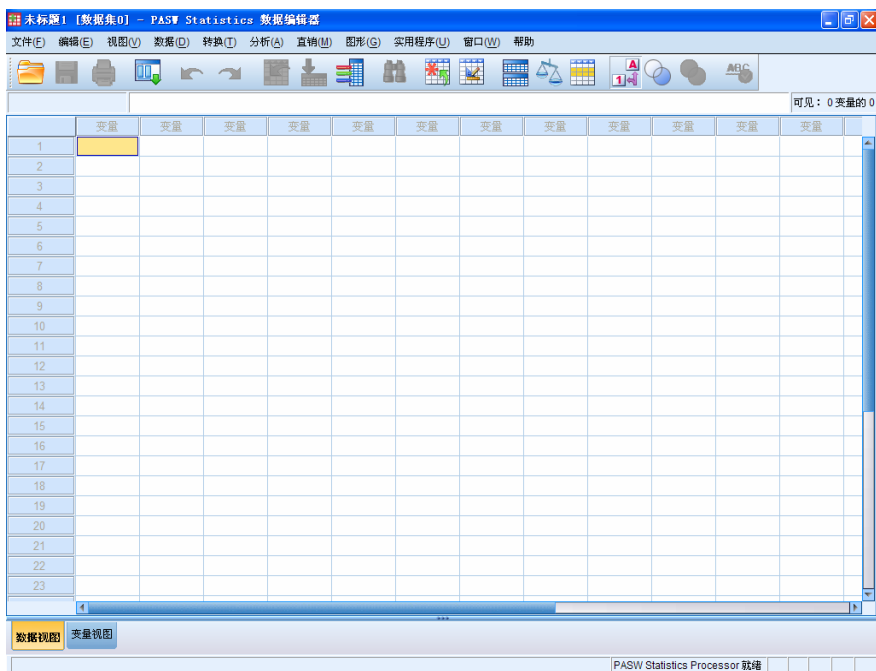


图 2-2 数据编辑器窗口

一些教材称数据的信息或者属性（例如变量名称、变量类型、存储类型、小数点的位数等）为数据字典。SPSS 变量视图用来定义 SPSS 数据集的数据字典。在数据编辑器窗口（左下角），选择“变量视图”标签，将出现如图 2-3 所示的变量视图。



图 2-3 变量视图

SPSS 的数据编辑器界面，无论是数据视图还是变量视图，都和 Excel 界面十分相近。数据视图界面包含窗口名显示栏、窗口控制按钮、SPSS 菜单、常用工具按钮、数据单元格信息显示栏、编辑显示区、编辑区选择栏、状态显示栏，具体说明如下。

窗口名显示栏：在窗口的顶部，显示窗口名称和编辑的数据文件名，没有文件名时显示为“未标题 1 [数据集 0]”。

窗口控制按钮：在窗口顶部的右上角，第一个按钮是窗口最小化，第二个按钮是窗口最大化，第三个按钮是关闭窗口。

SPSS 菜单：SPSS 菜单栏包含文件（File）、编辑（Edit）、视图（View）、数据（Data）、转换（Transform）、分析（Analyze）、直销（Direct Marketing）、图形（Graphs）、实用程序（Utilities）、窗口（Window）和帮助（Help）11 个子菜单。SPSS 英文版界面除了以上 11 个菜单以外，还有一个 Add-Ons 菜单，该菜单显示的是 SPSS 可以集成的 SPSS 的其他产品，例如 Amos、Modeler 等，如果购买了相关的产品，他们会出现在 SPSS 相应的菜单中。

数据单元格信息显示栏：和 Excel 类似，SPSS 工具栏的下方有单元格信息显示栏。在编辑显示区的上方，左边显示单元格和单元格所在列的变量名（单元格所在

行号：变量名），右边显示单元格里内容。

编辑显示区：在窗口的中部，最左边列显示数据行的序列号，最上边一行显示变量名称，默认为“变量”（英文界面默认为“Var”）。

编辑区选择栏：在编辑显示区下方，有两个标签，分别为数据视图（Data View）和变量视图（Variable View）。用户可以通过该选择栏在数据编辑区和变量编辑区之间进行切换。数据视图在编辑显示区中显示编辑数据，变量视图在编辑显示区中显示编辑数据的变量信息。

状态显示栏：在窗口的底部，左边显示执行的系统命令，右边显示窗口状态。


2. SPSS 的常用工具按钮


SPSS 菜单的下方是常用的工具按钮。在 SPSS 的数据视图下，在窗口显示的第三行上，SPSS 有 18 个工具按钮。这些工具按钮有：打开文档、保存文档、打印、对话检索、取消当前操作、重复操作、转到某条记录、转到某个变量、显示变量信息、查找、在当前记录的上方插入新的空白记录、在当前变量的左边插入新的空白变量、选择个案、拼写检查、分割文件、加权个案、使用变量集、显示值标签。如图 2-4 所示。




图 2-4 SPSS 工具栏


SPSS 有如下 18 个常见工具按钮。

：打开已经建立的数据文件。可以是 SPSS 的数据文件，*.sav 格式或者 .por（SPSS 便携数据）格式；或者从其他系统生成的数据文件（例如 SAS 数据文件、Excel 数据文件，Stata 数据文件等）。

：保存当前的数据文件。

：打印当前的数据文件。

：检索最近使用的对话框，可以快速地回到近期的对话框窗口，重复运行或者修改以前的分析程序对话框中的设置。

：取消用户最近的操作。



: 重复用户最近的操作。



: 转到某条记录或者个案。在大数据集文件中, 该工具按钮可以使用户快速地定位到指定的记录。



: 转到某个变量。该按钮可以快速地定位到指定的数据列。



: 显示变量信息。和【实用程序】菜单中的【变量】子菜单一样, 显示变量的数据字典。



: 在当前列 (或者某个变量) 中查找指定的值。



: 插入个案。该按钮将在当前位置插入一条空记录 (即一空行)。



: 插入变量。该按钮将在当前位置插入一新的列。



: 选择个案。和【数据】菜单中的【选择个案】子菜单的作用一样, 选择指定条件的记录。



: 拼写检查。和【实用程序】的子菜单【拼写】的作用一样, 用来检查变量标签和值标签中的拼写错误。



: 分割文件 (或者称为“拆分文件”)。按照分组变量将数据文件分割为单独的组, 然后根据一个或多个分组变量的值进行分析。



: 加权个案。用指定的频率变量对数据记录进行加权。



: 使用变量集。将数据编辑器和对话框变量列表中显示的变量限制为所选中的变量集合中的变量。



: 选择在数据视图中显示变量的值标签还是变量值 (即编码)。

3. 变量视图

变量视图可以定义和显示以下十一个变量属性。

1) 变量的名称: 给出变量或者属性的名称。变量名称需要符合 SPSS 变量名的命名准则:

- 必须以英文字母开头, 其他部分可以含有字母、数字、下画线 (即 “_”);
- 变量名尽量避免和 SPSS 已有的关键字重复, 例如 sum、compute、anova 等;

- SPSS13 及以后版本支持变量名最长为 64Byte，即变量名最长为 64 个英文字符，或者 32 个中文字符；
- SPSS 变量名不区分大小写，即 SPSS 认为 Name、name、nAme 这三个变量名没有区别。

2) 变量类型：选择变量的显示方式，如图 2-5 所示。

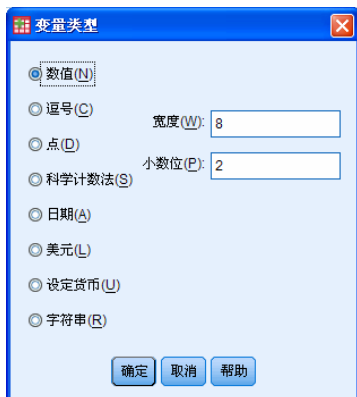


图 2-5 变量类型选择

- 数值型 (Numeric)：即常见的尺度 (Scale) 变量，或者经过编码的分类变量 (包含名义变量 (Nominal) 和有序变量 (Ordinal))。需要定义数据的总的宽度和小数的位数。选中“数值 (N)”以后，默认的数字宽度为 8，小数位为 2。可以根据需要进行修改。
- 逗号 (Comma)：即整数部分用逗号分隔的数值。在整数部分，从个位算起，每三位数一个逗号。小数点仍然为“.”。
- 点：即整数部分用点分隔的数值。在整数部分，从个位算起，每三位数用一个点分隔。小数点为“，”。
- 科学计数法 (Scientific)：用科学计数法来表示数值型数据。
- 日期 (Date)：日期型数据。
- 美元 (Dollar)：数据前有美元符号。可以选择具体数据的呈现方式。
- 设定货币：选用客户设定的货币格式。在应用该选项前，需要预先在选项中设置。方法为选择【编辑】→【选项】，进入选项设置对话框，然后选择“货币”标签，设定需要的货币格式，否则该格式不起作用。
- 字符串：如果变量或者属性是字符型数据，例如“性别”变量的取值为“男”和“女”，则我们须定义“性别”为字符串型。

3) 变量宽度：对字符型变量，该数值决定了你能输入的字符串的长度。

4) 小数位的宽度：设定小数位的宽度。

- 5) 变量标签: 给变量更详细的说明或描述。在分析过程和结果显示中, 可以选择显示变量名或者变量标签。
- 6) 变量的取值编码: 对变量值进行编码。
- 7) 缺失值编码: 对数据中的缺失值进行编码。
- 8) 列: 设定该变量数据视图中列的宽度。
- 9) 对齐方式: 列数据的对齐方式。
- 10) 变量度量类型: 设定变量度量标准, 有度量 (Scale)、序号 (Ordinal)、名义 (Nominal) 三种选择。度量型变量的数值有具体的度量意义, 例如个数、高度、温度等。序号型变量为分类变量中的有序变量, 比如编码的“十分重要”、“重要”、“一般”、“不重要”; 成绩中的 A、B、C 等都是序号变量。名义型变量为分类变量, 例如名字、地址、电话等。
- 11) 变量角色: 这是从 SPSS 版本 18 开始引入的 SPSS 数据挖掘软件 Modeler 中的一个数据属性, 用来指定该变量在建模中的角色是输入、目标或者不进入建模等。SPSS 18 以前的版本没有该变量属性。

注意: 由于 SPSS 不同的统计分析过程需要不同的数据类型, 因此, 在学习使用 SPSS 软件作统计分析时要注意变量的度量类型。变量的度量类型不是固定不变的, 可以根据分析过程来改变变量的度量类型。

2.3 新建数据文件、数据字典


刚刚完成一项调查或者试验, 可以把数据直接输入到 SPSS 软件中, 建立 SPSS 数据文件。一个好的习惯是, 在把数据输入 SPSS 以前, 先定义数据文件的结构。这要求先了解数据的构成, 每条记录有几个属性 (或变量), 每个属性的名称是什么, 这个属性是分类型数据还是连续型数据。这些清楚以后, 可以先进入变量视图, 把相应的变量定义好, 也就是等于把数据输入的模板定义好了, 然后到数据视图中输入数据。另外一种方式是, 先进入到数据视图, 采用 SPSS 默认的数据变量信息, 把数据先录入, 然后再到变量视图对变量属性进行相应的修改。

注意: SPSS 数据文件格式以每一行为一个记录, 或称观察单位 (Cases, 许多 SPSS 书籍翻译为“个案”); 每一列为一个变量 (Variable)。

现在，我们通过一个例子来学习数据的输入操作。我们对 12 个参加减肥活动的人做了一项调查。每个被调查者有一个 ID，然后调查他们的身高、参加活动以前的体重、参加活动以后的体重、性别、政治派别以及 8 个有关性格的问题。我们把这些收集到的调查问卷结果输入 SPSS 中，这些问题相对应的 SPSS 变量为：

- ID 号 (id)；
- 性别 (sex)；
- 身高 (height)；
- 参加活动以前的体重 (before)；
- 参加活动以后的体重 (after)；
- 政治派别 (party)；
- 8 个有关性格的问题（分别记为 e1 到 e8）。

打开 SPSS 数据编辑器的变量视图，按照下列步骤进行操作。

1. 在“名称”栏输入变量名“id”，单击类型栏中的“数值 (N)”单元格，该单元格变为带有省略号的图标（即 **数值(N)** ），单击该图标右侧的省略号，得到变量选择对话框，如图 2-6 所示。我们这里的变量“id”为数值型，因此，选择“数值”，把宽度改为 3，小数位设为 0，如图 2-6 所示。在标签栏，输入“问卷编号”，如图 2-6 所示。其他设置保持默认值。然后转到数据视图，在变量名“id”栏依次输入 1 到 12。

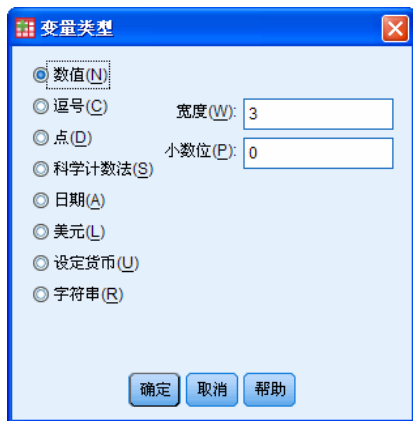


图 2-6 变量类型

2. 双击变量名“id”右边的变量或者单击“变量视图”标签，转到变量视图。在第二行输入“sex”，定义为数值型，宽度设为 1，小数位设为 0。在标签栏对应的格子中输入“性别”，在“值”栏对应的格子内，单击右侧带有省略号的图标

输入：值(U)：1，标签(L)：男。单击【添加】按钮，类似输入“女”。得到如图 2-7 所示的对话框。



图 2-7 值标签

其他设置保留默认值。转到数据视图，依次输入数据：1, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2。

3. 双击变量名 sex 右边的变量或者单击“变量视图”标签，转到变量视图。在第三行输入 before，定义为数值型，宽度设为 3，小数位设为 0。其他设置保持默认值。转到数据视图，依次输入数据：76, 59, 67, 65, 63, 72, 70, 68, 69, 74, 68, 63。类似地，输入变量 after, party, e1 到 e8。

数据输入完毕之后，变量视图如图 2-8 所示，数据视图如图 2-9 所示。

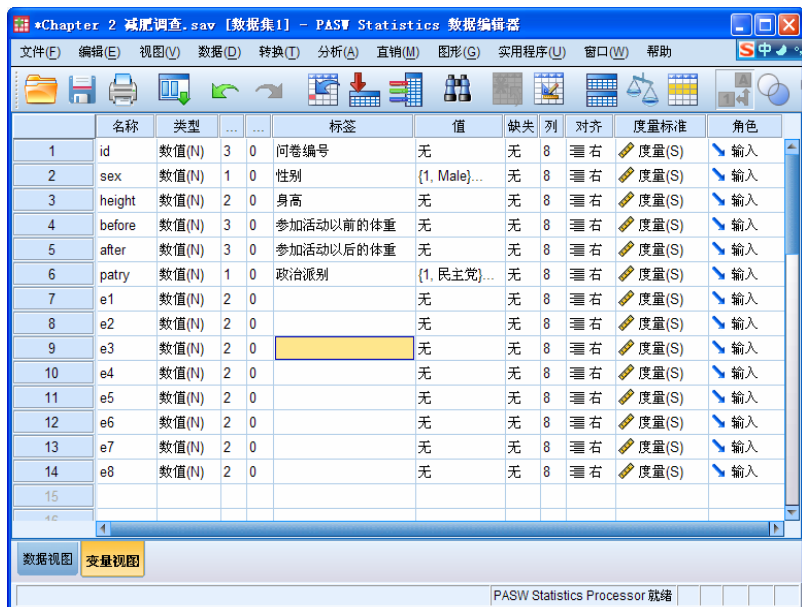


图 2-8 完成输入后的变量视图

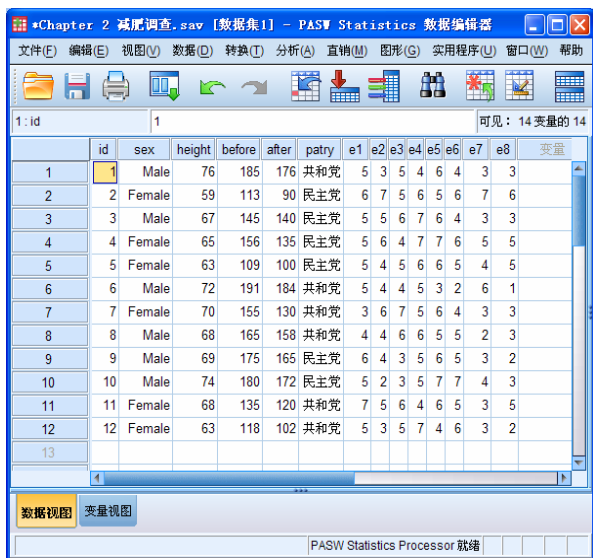


图 2-9 完成输入后的数据视图

SPSS 输入数据时候应该注意的问题。

1. 字符型数据

- 在 SPSS 中, 字符型数据值是区分大小写的, 小写的 m 和大写的 M 是不一样的。
- 字符型数据也可以设置值标签。例如, “sex” 变量的两个取值为 M、F, 他们的值标签分别为 Female、Male。
- 如果值标签为英语, 可以单击“拼写”进行拼写检查, 以检查值标签的英文拼写。

2. 缺失值处理

如果有数据缺失, SPSS 对于字符型数据和数值型数据有不同的处理方式。对于数值型数据, 缺失值默认为“.”; 对于字符型数据, 系统默认值为空, 如果空字符串有意义, 那么需要在变量视图对缺失值进行定义。

2.4 保存文件

在数据输入过程中, 要经常注意保存数据, 而不要等到所有数据输入完成之后再保存。这样可以避免不必要的数据丢失, 例如计算机故障或者突然断电造成的数据丢失。当单击【保存】按钮时, SPSS 可以对变量有选择地进行保存, 如图 2-10 所示。在保存数据对话框中, 当单击【变量】按钮时, 可以选择想要保存的变量, 默认保存数据文件中所有的变量。



图 2-10 保存数据对话框

在“文件名(N):”栏输入“Chapter 2 减肥调查”，然后单击【保存】按钮，该文件将被保存到在【选项】中所设置的工作目录下的“Chapter 2 减肥茶调查.sav”文件。在进行变量保存前，如果希望每次都在某个固定的目录下工作，可以设置 SPSS 工作目录。选择【编辑】→【选项】，打开如图 2-11 所示的“选项对”话框，选择“文件位置”标签，在“指定文件夹”部分，设置数据文件的目录和其他文件的目录（输出文件，语法文件等）。数据文件设置数据文件的默认位置，“其他文件”设置语法文件、结果文件等的默认位置。设置完成之后，每次打开或者保存文件时将指向这里所设定的目录。



图 2-11 “选项对”话框

2.5 读入数据

在 SPSS 文件菜单下，选择【文件】→【打开】→【数据】或者直接单击工具按钮栏上的“打开”按钮，将得到如图 2-12 所示的“打开数据”对话框。单击“文件类型（T）”右侧的向下箭头，列表给出了 SPSS 可以读入的数据文件类型。SPSS 特有的数据文件格式是后缀为.sav 的文件，或者后缀为.por 的文件。其他类型的数据格式有老版本的 SPSS 生成的数据（即 SPSS/PC+数据，后缀为.sys）、Systat 数据、SAS 数据、Stata 数据、Excel 表格、文本格式的数据等。根据需要打开的数据文件类型，在这里选择相应的文件类型。

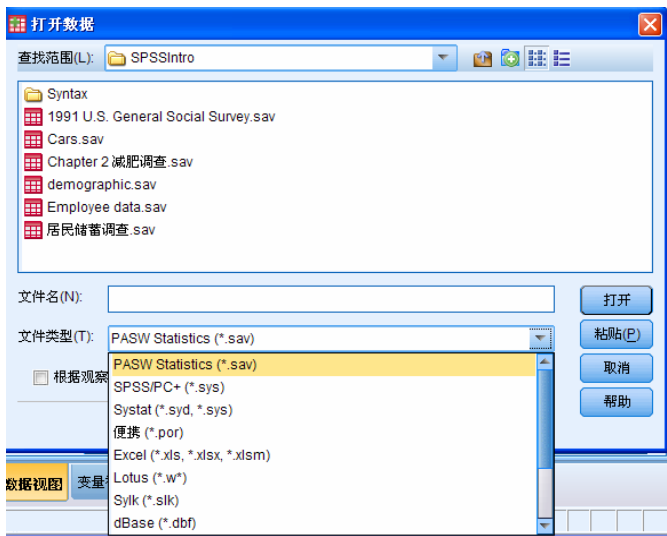


图 2-12 “打开数据”对话框

2.5.1 读入 Excel 数据

选择【文件】→【打开】→【数据】，文件类型选择 Excel，然后双击“Chapter 2 GSS04S.xls”，或者选中文件名，单击打开。将弹出一个“打开 Excel 数据源”对话框，以选择需要打开的工作表以及数据的范围。范围用 Xm:Yn 格式指定，X 代表 Excel 表格中开始读入的第一个数据的列名，m 代表行号；Y 代表 Excel 表格中读入的最后一个数据的列名，n 代表其所在的行号。默认情况选择 Excel 工作簿的最后工作的工作表和全部数据。默认情况下，SPSS 从第一行数据读入变量名。如果数据的第一行是数据，不要勾选“从第一行数据读入变量名”。

这里要打开的 Excel 文件的第一行是变量名，因此保留默认设置。如图 2-13 所示。

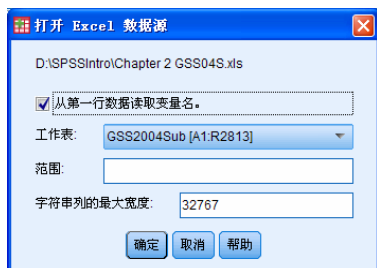


图 2-13 “打开 Excel 数据源”对话框

读入的数据在 Excel 2003 中的显示如图 2-14 所示。

	Gender	Age	Education	Fathers Education	Mothers Education	Spouses Education	Degree	Age First Kid	Race	Household Income	Response Income	Income Dollars	Region	Work Status	Hours Worked	Spouse
1	F	52	9	9	10	8	0	34	3	12	.	.	2	7	.	.
2	F	43	14	13	1	12	1	35	3	5	.	.	2	7	.	.
3	M	52	14	97	12	97	1	23	2	12	99.00	9999...	2	1	40.00	.
4	M	53	14	1	3	97	2	19	1	12	12.00	4403...	2	1	48.00	.
5	F	34	17	97	18	97	3	.	3	12	12.00	5572...	2	1	36.00	.
6	M	33	16	8	12	97	3	.	2	10	10.00	1889...	2	4	.	.
7	M	24	14	12	12	97	1	.	3	12	12.00	9779...	2	1	35.00	.
8	M	22	14	14	97	97	1	.	3	2	.	.	2	2	99.00	.
9	M	19	12	14	11	97	1	.	2	12	.	.	2	2	25.00	.
10	M	26	14	11	12	97	2	.	2	12	12.00	6362...	2	1	40.00	.
11	M	74	12	0	0	97	1	19	2	1	.	.	2	6	.	.
12	F	58	12	5	6	97	1	21	2	12	12.00	7096...	2	1	75.00	.
13	M	32	10	10	0	11	0	28	2	11	9.00	1449...	2	1	45.00	.
14	M	40	18	8	6	17	4	19	2	12	12.00	3743...	2	2	12.00	.
15	M	42	14	97	16	97	2	.	2	12	11.00	2174...	2	1	40.00	.
16	M	48	11	97	98	97	0	25	2	2	3.00	3590...	2	7	.	.
17	F	39	9	98	6	97	0	18	2	13	3.00	3990...	2	1	40.00	.
18	M	27	9	97	10	97	0	.	1	10	10.00	1654...	2	1	40.00	.
19	F	38	16	10	14	14	3	25	2	13	.	.	2	1	40.00	.
20	M	29	11	12	12	97	1	25	3	12	12.00	6508...	2	2	30.00	.
21	F	29	13	10	16	97	0	16	2	98	.	.	2	1	50.00	.
22	F	67	14	97	8	12	1	21	1	12	.	.	2	1	99.00	.
23	M	67	12	16	12	12	1	21	1	0	0.00	1215...	2	4	.	.

图 2-14 Excel 源数据

- 注意：**
1. SPSS 只是读入数据，其他和 Excel 单元格关联的属性，例如注释、公式等，都不会被读入 SPSS 文件。因此，在读入 Excel 数据文件以前，需要确保 Excel 文件中含有的只是数据。在读入 SPSS 以前，需要先删除非数据部分单元格内的内容或者其他和需要处理的数据无关的部分的单元格的内容。
 2. 从 SPSS 16 开始，可以读入 Excel 2007 数据文件。SPSS 15 以及以前的版本不能够读取 Excel 2007 数据文件。
 3. 在 SPSS 读入 Excel 文件时，必须先关闭要读入的 Excel 数据文件，否则 SPSS 软件读取数据时会报错。
 4. 建议在读入 Excel 文件以前，先仔细检查 Excel 文件中的数据，确保格式正确，并删除和数据无关的部分以及空行和空列，然后再运行 SPSS 读入该文件。

2.5.2 读入文本数据

文本数据是最常见的数据格式之一。大部分的数据库和数据分析软件都可以把数据保存为文本格式。常见的文本数据有两种格式：分隔符分隔的数据文件和固定列宽的数据文件。本节将介绍 SPSS 读入分隔符分隔的数据文件的方法。

单击菜单【文件】→【打开】，文件类型选择“文本文件 (*.txt, *.dat)”。在“打开数据”窗口，选择“Chapter 2 GSS04S.txt”，如图 2-15 所示。

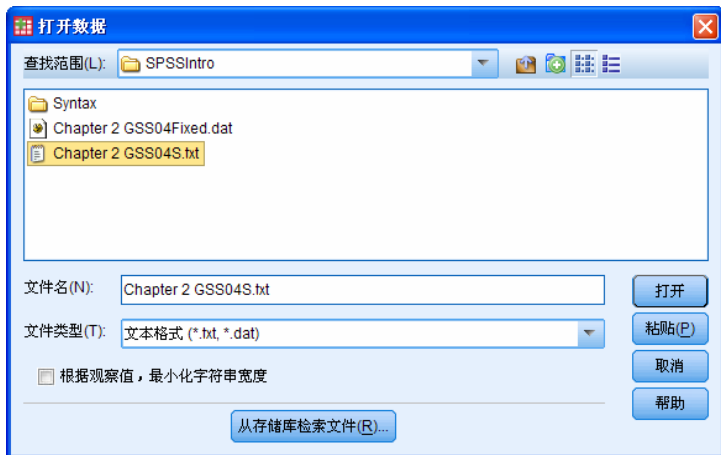


图 2-15 “打开数据”窗口

单击【打开】按钮，将出现“文本导入向导”对话框，该向导一共有六步，它将指导用户完成文本数据文件的导入任务，具体说明如下。

第 1 步：文本导入向导的第 1 步将显示数据预览。从预览中可以知道 SPSS 按照默认方式读入的数据是否正确。这里，SPSS 默认第一行读入的数据是数据的表头，即变量名。如果默认读入的方式是有错误的，可以在以后的第 2 至第 6 步中，对默认的设置进行修正。

注意：如果以前导入过相同格式的文本数据文件，并且把导入的格式保存下来，则可以在“您的文本文件与预定义的格式匹配吗？”框中选择“是”，然后单击“浏览”选择以前所预定义的格式文件，之后直接单击对话框下部的【完成】按钮，文件的导入操作即告完成，没有必要再进行第 2 至第 6 步的操作。

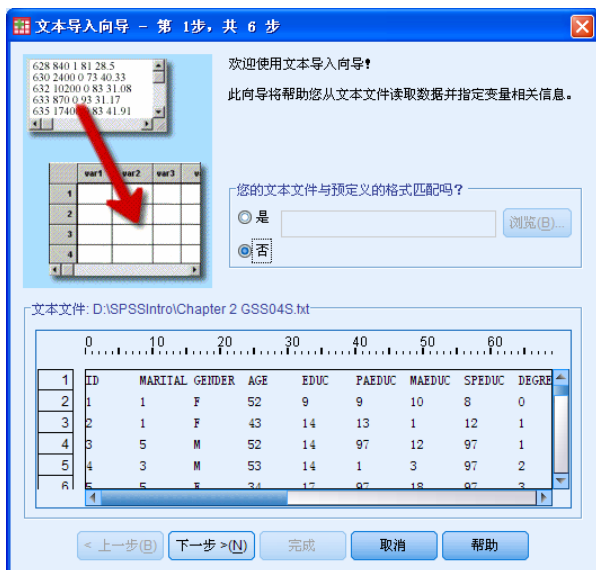


图 2-16 文本导入向导第 1 步

第 2 步：设置文本文件中变量的排列方式和变量名的设置。

这里，要读入的文本文件的变量是由分隔符（制表符）分隔的，并且文件的顶部是变量名称。因此，在该对话框中做如图 2-17 所示的设置。



图 2-17 文本导入向导第 2 步

第 3 步：设定从何处开始导入个案和导入个案的个数。

文本导入向导的第 3 步将设置数据开始的位置、个案的分隔方式以及需要导入的个案的个数，并且呈现采用设定后数据的预览。本例中的设置如图 2-18 所示：



图 2-18 文本导入向导第 3 步

第 4 步：设定个案内变量的分隔方式。在“变量之间有哪些分隔符?”对话框中，有五种选择，分别是：制表符、空格、逗号、分号、其他。根据文本文件的格式，选定相应的分隔符。在“文本限定符是什么?”部分用来选定用什么方式来标识文本或者字符串，根据文本数据文件中的具体情况来设定“无”（即没有特殊标识）、“单引号”、“双引号”或者“其他”。该对话框的“数据预览”部分显示按照设定情况所读入的数据的预览，如图 2-19 所示。



图 2-19 文本导入向导第 4 步

从图 2-19 的预览看出许多列没有数值，并且多出了 V19 到 V22 四列，读入的

最后四列不正确，因此怀疑按照默认方式读入数据是有错误的。这里需要借助文本编辑器打开需要读入的数据文件，检查是否有多余的列。经检查，文本文件中的数据和数据标题对应，没有多余的列，出错原因是读入向导中的变量分隔符选择不正确，因而需要更改变量间的默认分隔符。这里，“变量之间有哪些分隔符？”部分只勾选“制表符（T）”，去掉“空格（S）”前面的钩，如图 2-20 所示。



图 2-20 文本导入向导第 5 步

注意：1. 一定要正确设定变量间的分隔符才能够保证导入数据的正确性。
2. 在单击【下一步】按钮之前，要仔细检查数据预览部分，如发现不正确的部分需要找到出错的原因，并及时返回相应的步骤进行修改。

第 5 步：对变量名以及数据格式进行调整。SPSS 读入文本数据时，会根据指定的位置读入变量名，如果指定的变量名不符合 SPSS 变量名的命名规则，SPSS 导入时会进行转换。例如，如果指定的变量名中含有空格，则 SPSS 会自动地把空格删除。

单击“数据预览”部分变量名，可以调整变量名，或者检查变量的数据格式。在这里，AGE 的数据格式读入时默认为字符串，我们改为数值；另外，INCOME_ACTUAL 默认的格式为数值，我们改为“美元”。如图 2-21 和图 2-22 所示。



图 2-21 变量名称和数据格式预览



图 2-22 更改变量名称和数据格式

第 6 步：指定是否保存该导入数据的方式为文件格式并将其保存为语法文件。这里我们不保存该格式，即按照默认的方式选择“否”，如图 2-23 所示。

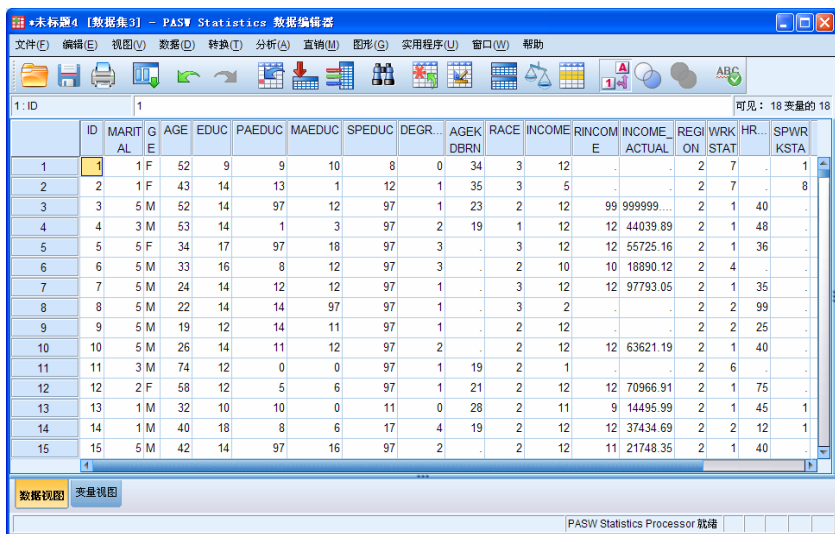


图 2-23 文本导入向导第 6 步

单击【完成】按钮，最后在 SPSS 数据视图中得到导入的数据如图 2-24 所示。

注意：导入文本数据时，一定要在每一个步骤中检查数据预览，确定数据预览符合预期的格式；如果不正确的数据格式出现，回到相应的步骤调整相应的设置，直到数据预览正确为止。

如果尝试了所有可能的设置，读入数据仍然不正确，需要仔细检查原始的文本数据，确保原始的文本数据文件格式正确。



	ID	MARITAL	GENDER	AGE	EDUC	PAEDUC	MAEDUC	SPEDUC	DEGR	AGEK	RACE	INCOME	RINCOME	INCOME	REG	WRK	HR	SPWR
1	1	1	F	52	9	9	10	8	0	34	3	12			2	7		1
2	2	1	F	43	14	13	1	12	1	35	3	5			2	7		8
3	3	5	M	52	14	97	12	97	1	23	2	12	99	999999	2	1	40	
4	4	3	M	53	14	1	3	97	2	19	1	12	12	44039.89	2	1	48	
5	5	5	F	34	17	97	18	97	3		3	12	12	55725.16	2	1	36	
6	6	5	M	33	16	8	12	97	3		2	10	10	18890.12	2	4		
7	7	5	M	24	14	12	12	97	1		3	12	12	97793.05	2	1	35	
8	8	5	M	22	14	14	97	97	1		3	2			2	2	99	
9	9	5	M	19	12	14	11	97	1		2	12			2	2	25	
10	10	5	M	26	14	11	12	97	2		2	12	12	63621.19	2	1	40	
11	11	3	M	74	12	0	0	97	1	19	2	1			2	6		
12	12	2	F	58	12	5	6	97	1	21	2	12	12	70966.91	2	1	75	
13	13	1	M	32	10	10	0	11	0	28	2	11	9	14495.99	2	1	45	1
14	14	1	M	40	18	8	6	17	4	19	2	12	12	37434.69	2	2	12	1
15	15	5	M	42	14	97	16	97	2		2	12	11	21748.35	2	1	40	

图 2-24 导入数据完成后的数据视图

2.5.3 读入数据库数据

SPSS 可以读入所有类型的数据库文件，例如 Access、SQL Server、Oracle、DB2、MySQL 等。所有的数据库文件都可以通过建立 ODBC 数据源的方式来读入到 SPSS 中。MS Access 和 MS Excel 类型的数据文件已在经 ODBC 数据源中列出，SPSS 可以直接读入 Excel 和 MS Access 数据库的数据文件。其他类型的数据库文件需要先建立 ODBC 数据源。

这里用 Access 数据库文件 Chapter 2 GSS04.mdb 为例来说明 SPSS 导入数据库文件的过程。Chapter 2 GSS04.mdb 数据库包含两张表：GSS2004Add 和 GSS2004Sub。其中，GSS2004Sub 包含被调查者的人口统计学信息，例如婚姻状况、年龄、性别、教育程度等；GSS2004Add 包含被调查者对主要调查问题的反馈，例如工作是否满意、对幸福情况的反馈，对健康状况的反馈等。这两张数据表在 Access 数据库中的视图，如图 2-25 所示。

SPSS 数据库向导遵循下列步骤。

步骤一：选择数据源

选择【文件】→【打开数据库】→【新建查询】，出现如下的“数据库向导”，选择 MS Access Database，如图 2-26 所示。

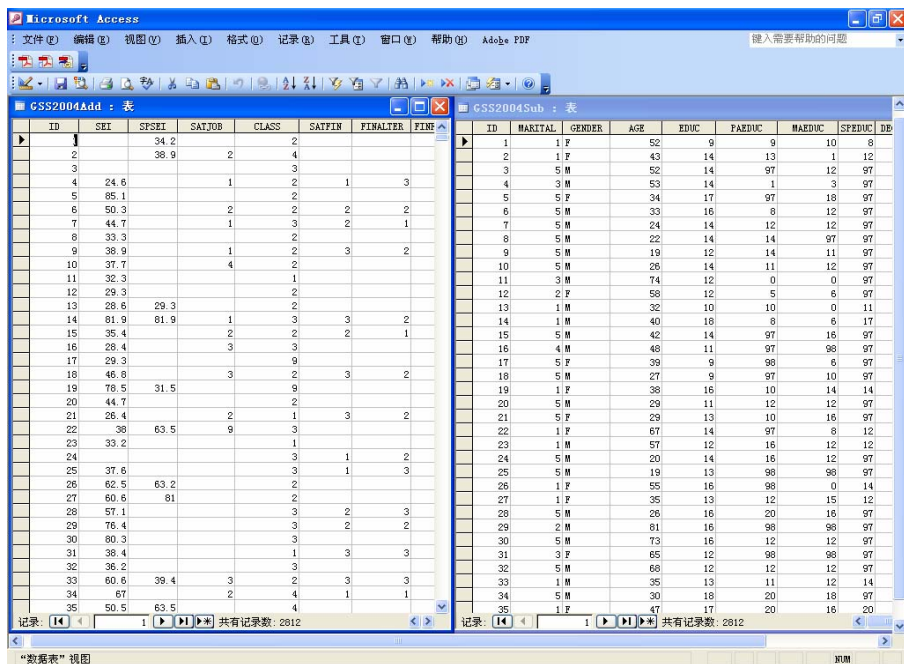


图 2-25 两张数据表在 Access 数据库视图



图 2-26 数据库向导——选择 ODBC 数据源

单击【下一步】按钮，将弹出“ODBC 驱动程序登录”窗口，单击【浏览】按钮，选择相应的数据库文件。这里选择 Chapter 2 GSS04.mdb，如图 2-27 所示。



图 2-27 选择数据库文件

步骤二：选择数据库表及其字段

单击【确定】按钮，将出现“数据库向导—选择数据”对话框，它将指导用户选择需要的字段（或者属性）。如图 2-28 所示，“可用表格（A）”中显示选定的数据库中的所有可用数据表，右栏显示选定的表格中的字段（即将被导入 SPSS 中的变量或字段）。这里选择两个数据表中的所有字段，如图 2-28 所示。

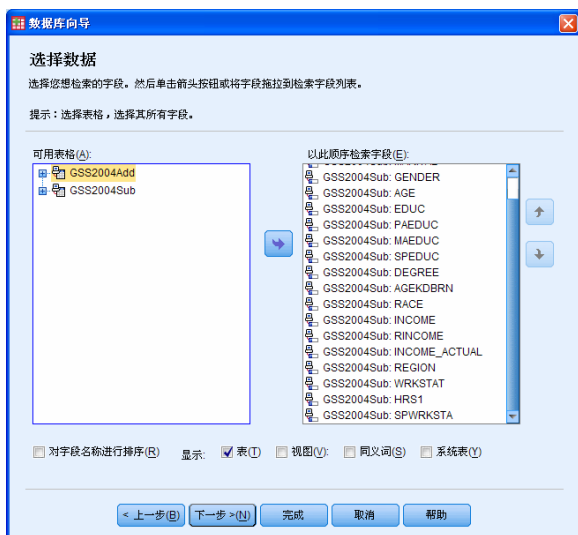


图 2-28 选择数据库表和字段

注意：1. SPSS 可以一次读入两个或者两个以上的表。读入多个表时，需要先把需要读入的表的字段选入右栏，然后单击【下一步】按钮指定连接类型。连接类型有三种：内部关联、外部右侧和外部左侧。

2. 双击左侧可用表格中的数据库表或者数据库表的某个字段将直接把该数据表的所有字段或者选定的字段选入右侧的检索字段中。

步骤三：选择多个数据表之间的连接关系

这里把 Access 数据库中的两个数据表：GSS2004Add 和 GSS2004Sub 都导入 SPSS 中。单击【下一步】按钮，采用默认关联类型——内部关联，两张数据表的 ID 作为默认关联关键字，它们之间有一个双箭头关联，如图 2-29 所示。

步骤四：限制检索个案

单击【下一步】按钮，出现限制检索个案向导，如图 2-30 所示。该对话框可以指定多个数据表之间的连接条件、单一数据表中选择个案的条件和数据的抽样方式。



图 2-29 指定两个表之间的关联关系

如果选中“随机抽样个案”，可以在满足条件的数据中随机选择一部分导入到 SPSS 中。抽样条件由采样方法和样本尺寸指定。

- 采样方法：在数据库中检索个案，在 SPSS Statistics 中随机选择或者在数据库中随机选择并在 SPSS Statistics 中检索。
- 样本尺寸：默认导入全部数据，也可以导入一定百分比的数据或者导入精确个数的个案。

注意：对于大数据源，可能需要将个案限制为数量较少的、具有代表性的样本，这可以显著减少其运行程序所需的时间。限制检索个案的界面如图 2-30 所示。



图 2-30 限制检索的个案

在图 2-30 中，保留默认设置。

步骤五：变量属性的调整

以上所有步骤定义完成后，SPSS 数据库向导会给出将要导入 SPSS 中的变量及其属性的列表，如图 2-31 所示。如果数据库中字段的数据类型为字符串，在这一步可以选择重新编码为数值，导入的相应数据在 SPSS 中会自动进行编码，原来的字符串将作为值标签。



图 2-31 编辑变量名和变量属性调整

步骤六：生成 SQL 语法

步骤六是最后一步，生成以上数据库查询的 SQL 语法。如果需要，可以保存下来以备在其他程序中使用，如图 2-32 所示。我们这里不保存 SQL 查询语句，保留默认的设置。



图 2-32 数据库导入的 SQL 语句

单击【完成】按钮，读入的数据在 SPSS 中的视图如图 2-33 所示。

	ID	MARITAL	GENDER	AGE	EDUC	PAEDUC	MAEDUC	SPEED	DEGREE	AGEKID	RACE	INCOME	RINCOME	INCOME	REGISTRATION	WRSAT	HRS1	SPW	ID1	SEI	SPS	SAT	JOB
1	1.00	1.00	F	52.00	9.00	9.00	10.00	8.00	.0	34.00	3.00	12.00	.	.	2.00	7.00	.	1.00	1.00	.	34.20	.	.
2	2.00	1.00	F	43.00	14.00	13.00	1.00	12.00	1.00	35.00	3.00	5.00	.	.	2.00	7.00	.	8.00	2.00	.	38.90	2.00	.
3	3.00	5.00	M	52.00	14.00	97.00	12.00	97.00	1.00	23.00	2.00	12.00	99.00	9999...	2.00	1.00	.	40.00	3.00
4	4.00	3.00	M	53.00	14.00	1.00	3.00	97.00	2.00	19.00	1.00	12.00	12.00	4403...	2.00	1.00	.	48.00	4.00	24.60	.	1.00	.
5	5.00	5.00	F	34.00	17.00	97.00	18.00	97.00	3.00	.	3.00	12.00	12.00	5572...	2.00	1.00	.	36.00	5.00	85.10	.	.	.
6	6.00	5.00	M	33.00	16.00	8.00	12.00	97.00	3.00	.	2.00	10.00	10.00	1889...	2.00	4.00	.	.	6.00	50.30	.	2.00	.
7	7.00	5.00	M	24.00	14.00	12.00	12.00	97.00	1.00	.	3.00	12.00	12.00	9779...	2.00	1.00	.	35.00	7.00	44.70	.	1.00	.
8	8.00	5.00	M	22.00	14.00	14.00	97.00	97.00	1.00	.	3.00	2.00	.	.	2.00	2.00	.	99.00	8.00	33.30	.	.	.
9	9.00	5.00	M	19.00	12.00	14.00	11.00	97.00	1.00	.	2.00	12.00	.	.	2.00	2.00	.	25.00	9.00	38.90	.	1.00	.
10	10.00	5.00	M	26.00	14.00	11.00	12.00	97.00	2.00	.	2.00	12.00	12.00	6362...	2.00	1.00	.	40.00	10.00	37.70	.	4.00	.
11	11.00	3.00	M	74.00	12.00	.0	.0	97.00	1.00	19.00	2.00	1.00	.	.	2.00	6.00	.	.	11.00	32.30	.	.	.
12	12.00	2.00	F	58.00	12.00	5.00	6.00	97.00	1.00	21.00	2.00	12.00	12.00	7096...	2.00	1.00	.	75.00	12.00	29.30	.	.	.
13	13.00	1.00	M	32.00	10.00	10.00	.0	11.00	.0	28.00	2.00	11.00	9.00	1449...	2.00	1.00	.	45.00	1.00	13.00	28.60	29.30	.
14	14.00	1.00	M	40.00	18.00	8.00	6.00	17.00	4.00	19.00	2.00	12.00	12.00	3743...	2.00	2.00	.	12.00	1.00	14.00	81.90	81.90	1.00
15	15.00	5.00	M	42.00	14.00	97.00	16.00	97.00	2.00	.	2.00	12.00	11.00	2174...	2.00	1.00	.	40.00	15.00	35.40	.	2.00	.
16	16.00	4.00	M	48.00	11.00	97.00	98.00	97.00	.0	25.00	2.00	2.00	3.00	3590...	2.00	7.00	.	.	16.00	28.40	.	3.00	.
17	17.00	5.00	F	39.00	9.00	98.00	6.00	97.00	.0	18.00	2.00	13.00	3.00	3990...	2.00	1.00	.	40.00	17.00	29.30	.	.	.
18	18.00	5.00	M	27.00	9.00	97.00	10.00	97.00	.0	.	1.00	10.00	10.00	1654...	2.00	1.00	.	40.00	18.00	46.80	.	3.00	.
19	19.00	1.00	F	38.00	16.00	10.00	14.00	14.00	3.00	25.00	2.00	13.00	.	.	2.00	1.00	.	40.00	1.00	19.00	78.50	31.50	.
20	20.00	5.00	M	29.00	11.00	12.00	12.00	97.00	1.00	25.00	3.00	12.00	12.00	6508...	2.00	2.00	.	30.00	20.00	44.70	.	.	.
21	21.00	5.00	F	29.00	13.00	10.00	16.00	97.00	.0	16.00	2.00	98.00	.	.	2.00	1.00	.	50.00	21.00	26.40	.	2.00	.
22	22.00	1.00	F	67.00	14.00	97.00	8.00	12.00	1.00	21.00	1.00	12.00	.	.	2.00	1.00	.	99.00	5.00	22.00	38.00	63.50	9.00

图 2-33 将数据库中数据导入到 SPSS 数据视图中

注意：读入两个或者两个以上的数据库表，必须指定数据库间的关联方式。大部分情况下可以选择默认“内部关联”，该种关联确保完全匹配关联条件的记录才会被导入 SPSS 中。如果采用外部左侧，则双箭头左侧的数据表中即使和右侧的数据表中不匹配的记录也会被选入 SPSS 数据集中。采用外部右侧，则双箭头右侧的数据表中即使和左侧的数据表中不匹配的记录也会被选入 SPSS 数据集中。

2.6 数据文件的合并

有时候，要把多个数据文件合并为一个数据文件。例如，一个公司在全国各地有 30 多个分公司，每个月公司总部需要把各分公司的销售人员的销售情况合并到一个数据文件中。又例如，一个学校教务部门每一个学期对学生的成绩建立一个数据文件，在最后学生毕业前，需要把学生四个学年（共 8 个学期）的成绩进行汇总，这就需要把 8 个数据文件合并为一个数据文件。

上述第二个例子中的合并 8 个成绩文件和第一个例子中的合并 30 个子公司的销售情况是不同的。学生的 8 个学期的成绩文件中，除了学生姓名不变以外，其他变量（或者属性）是不同的，每个学期的学习科目是不同的。而第一个例子中的每个子公司销售情况数据文件中的变量（或者属性）都相同，不同的是销售人员。

针对以上两种不同的情况，合并数据文件分为添加变量（或称为合并变量）和添加个案（或称为合并记录、合并个案）。在 SPSS 中的操作方式分别为添加个案和添加变量两种。

1) 添加个案：选择【数据】→【合并文件】→【添加个案】，如图 2-34 所示。

2) 添加变量：选择【数据】→【合并文件】→【添加变量】，也如图 2-34 所示。

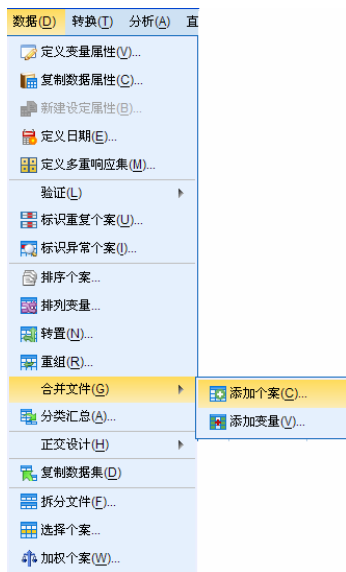


图 2-34 合并数据文件的菜单

下面举例说明这两种 SPSS 合并数据文件的方法。

2.6.1 添加个案

1. 合并两个数据文件

当进行合并个案时，两个数据文件变量的顺序、变量名称和变量个数可以不同。如图 2-35 所示，两个销售文件分别有共同属性（或变量）销售 ID、年龄、性别、销售额。但是，两个文件变量的顺序不同，且同一个属性在两个文件中的名称不同：子公司 1 的销售文件命名为“销售额”，而子公司 2 的销售文件命名为“销售金额”。子公司 1 和子公司 2 的销售文件有两个不同变量：职务和销售费用。由于子公司 1 的销售文件没有“销售费用”变量，合并后的文件中对应于子公司 1 的个案的“销售费用”自动取为系统默认缺失值（“.”）。同理，对应于子公司 2 个案的“职务”变量自动取值为系统默认缺失值（这里为空字符串）。

子公司1销售文件					子公司2销售文件				
销售ID	年龄	性别	销售额	职务	销售ID	性别	销售金额	年龄	销售费用
1	33	男	129.9	销售	4	男	187.2	28	1.2
2	36	女	300.8	销售经理	5	男	376.1	32	2.2
3	35	男	412.2	销售	6	女	432.1	38	2.9
					7	男	421.6	37	2.1
合并后的文件（添加案例）									
销售ID	年龄	性别	销售额	职务	销售费用				
1	33	男	129.9	销售	.				
2	36	女	300.8	销售经理	.				
3	35	男	412.2	销售	.				
4	28	男	187.2		1.2				
5	32	男	376.1		2.2				
6	38	女	432.1		2.9				
7	37	男	421.6		2.1				

图 2-35 添加个案示例图

打开两个数据文件 Sales1.sav 和 Sales1.sav，选择 Sales1.sav 为当前工作数据文件。选择【数据】→【合并文件】→【添加个案】，得到如图 2-36 所示的添加个案步骤 1 对话框。

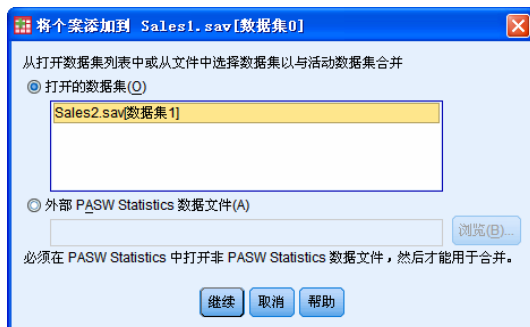


图 2-36 添加个案步骤 1

单击【继续】按钮，得到如图 2-37 所示添加个案步骤 2 对话框。

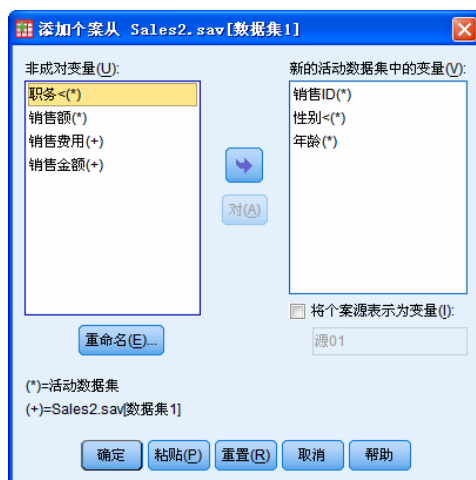


图 2-37 添加个案步骤 2

SPSS 用“（*）”表示变量来自于当前活动数据文件中的变量，而用“（+）”表

示将要和当前数据文件进行合并的数据文件中的变量。图 2-37 右栏中表示两个数据文件中都有的变量，如果把该栏中的变量移到左栏，则该变量将不会在合并后的文件中出现。图 2-37 左栏中表示两个数据文件中不能匹配的变量，即在另一个文件中不能找到同名的变量。来自于 Sales1.sav 中的变量职务、销售额和来自于 Sales2.sav 中的销售费用、销售金额不能匹配。我们知道销售额和销售金额是同一属性，只是变量名称不同。我们首先选中“销售额”，然后在按住“Ctrl”键的同时选中“销售金额”，单击按钮【对(A)】，这样 SPSS 就自动把这两个变量作为同一变量，并在合并后的文件中命名为“销售额”。然后，同时选中左边框中剩余的变量，把它们选到右边的框中。如果选中图 2-38 右下角的“将个案源表示为变量(I)”，则会在合并后的文件中生成一个新的变量，用它来标识个案是来自于哪个数据文件。这里保留默认值（即选中该选项）。



图 2-38 添加个案步骤 3

单击【确定】按钮，得到的当前工作文件即为合并后的数据文件。这里，最后一列即为勾选了“将个案源表示为变量(I)”后生成的新变量，变量名称默认为“源01”。“0”表示个案来源于当前活动数据文件，即 Sales1.sav，“1”表示来源于“Sales2.sav”。

SPSS 并没有为合并后的数据文件生成一个新的数据文件，而是把 Sales2.sav 的个案直接添加到当前工作文件即 Sales1.sav 中。如果之后还会用到第一个数据文件，需要把合并后的数据文件另存为不同于当前工作文件的文件名，这里另存为 Sales1_Merge_Sales2.sav，如图 2-39 所示。

1. 销售ID	年龄	性别	销售金额	职务	销售费用	源01	变量	变量	变量	变量	变量	变量
1	h	33 男	130	销售	.	0						
2	2	36 女	301	销售	.	0						
3	3	35 男	412	销售	.	0						
4	4	28 男	187		1.20	1						
5	5	32 男	376		2.20	1						
6	6	38 女	432		2.90	1						
7	7	37 男	422		2.10	1						
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												

图 2-39 合并个案后的数据文件

2. 合并两个以上数据文件

如果有三个或者以上的数据文件需要合并，可以先合并两个，依次进行。如果需要合并的文件个数较多，这种方式比较费时。是否能一次合并三个或者以上的数据文件呢？答案是肯定的。我们可以通过编写 SPSS 语法程序的方式来实现，该方式可以一次性合并 2 到 50 个数据文件。假设我们要合并四个数据文件，他们分别为 Sales1.sav、Sales2.sav、Sales3.sav、Sales4.sav。我们可以先打开 Sales1.sav 作为当前工作数据文件，然后运行下列命令：

```
ADD FILES /FILE=*
/FILE='C:\SPSSIntro\Chapter 2\Sales2.sav'
/RENAME 销售金额=销售额
/FILE='C:\SPSSIntro\Chapter 2\Sales3.sav'
/FILE='C:\SPSSIntro\Chapter 2\Sales4.sav'.
EXECUTE.
```

最后得到的当前工作文件即为合并四个文件后的文件。

注意：1.SPSS 可以同时打开多个数据文件，原则上可以用复制粘贴的方式来合并这些文件。这种方式是不推荐使用的。如果这些文件变量的顺序不一致或者变量并不完全一样，用这种方式合并的文件是错误的。另外，也不会有标识个案来源的新变量生成。

2.合并两个数据文件时，SPSS 并没有生成一个新的数据文件，而是把被合并文件的个案直接添加到当前工作文件中。如果之后还会用到第一个数据文

件，需要选择“另存为”把合并后的数据文件另存为不同于当前工作文件的文件名。

3. 当文件合并完成之后，一定要检查合并后的数据文件是否正确。例如，可以应用探索性统计分析方法来检查合并后的文件。
4. 需要合并的数据文件可以是 SPSS Statistics 中已经打开的数据文件，或者没有打开的 SPSS Statistics 格式的数据文件(*.sav 格式)，也可以是其他格式的数据文件，如.dBase 文件等。

2.6.2 添加变量

如果有两个数据文件，它们含有相同的个案，但是不同文件含有的属性不同，现在需要把这两个文件合并为一个文件。合并这样的数据文件就是添加变量。如果需要合并的数据文件中都含有同一个变量，该变量可以用于标识这些数据文件中的个案，并且可以按照该变量来匹配这些数据文件中的记录，该变量称为关键变量。例如在应用一种新的分发系统之前对客户做过调查，数据保存在一个文件中；应用该系统之后又对以前的客户做了相同的调查，数据保存在另一个数据文件中。我们想知道哪些客户会受益于新的分发系统，那么首先要根据关键变量（即客户编号）把这两个数据文件合并在一起。又比如，学生的成绩按照学期分别存在八个数据文件中，毕业之前需要对学生的成绩做出综合评价，这就需要先把这八个数据文件按照学生姓名（或者学生 ID）合并在一起。

添加变量合并文件有两种情况：一对一合并和一对多合并。一对一合并时，两个数据文件是对等的，两个文件中的个案之间是一一对应的，否则取系统缺失值。而一对多合并时，一个文件会作为主文件，或者称为查表文件，该文件中的一个个案可以和另一个文件中的多个个案相匹配。

以添加变量方式合并两个或者多个数据文件时，需要注意下列问题：

- 在合并数据文件之前，所有需要合并的数据文件必须预先按照关键变量进行升序排列。否则，合并文件程序将失败；
- 与添加个案不同，这里所有需要合并的数据文件必须是 SPSS Statistics 定义的数据文件格式(.sav 格式)或者已经在 SPSS Statistics 中打开的数据文件；
- 由于一个文件中的变量需要添加到另一个文件，必须确保两个文件中需要合并的变量名称不同。

1. 一对一合并

- 对于两个数据文件，如果进行一对一合并，一个文件中的每个个案只能根据关键变量匹配另一个文件中唯一的个案。反之亦然。
- 如果一个文件中的某个个案在另一个文件中找不到个案来匹配，则该个案于第二个文件的变量上的取值为缺失值。反之亦然。
- 如果一个文件中的某个个案在另一个文件中找到两个或者两个以上的个案来匹配，则该个案只取第二个文件中第一个相匹配的个案来连接。反之亦然。

假设我们有两个数据文件，一个称为 1960 年数据文件，记录了 1960 年各个国家的人口数据；另一个称为 1990 年数据文件，记录了 1990 年各个国家的人口数据。我们把两个数据文件合并在一起，比较各个国家 30 年来人口的变化情况。

如图 2-40 所示，国家作为关键变量。第一个文件的 Ghana 所在的个案在 1990 年数据文件中找不到匹配的国家，因此合并后的数据文件中 Ghana 在第二个数据文件的人口变量上为缺失值。在 1990 年数据文件中，有两个个案和 1960 年的个案 Russia 匹配，合并后的数据文件只取第一个匹配的个案。

1960 年数据文件		合并后的文件			1990 年数据文件	
国家	人口	国家	60 年人口	90 年人口	国家	人口
Austria	X	Austria	X	Y	Austria	Y
Ghana	X	Ghana	X	.		.
Russia	X	Russia	X	Y	Russia	Y
.		Russia	.	Y	Russia	Y
US	X	US	X	Y	US	Y
		Yemen	.	Y	Yemen	Y

图 2-40 一对一添加变量

在 SPSS 中，添加变量是通过菜单【数据】→【合并文件】→【添加变量】来完成的（如图 2-34 所示）。打开 SPSS 的两个数据文件：World60.sav 和 World90.sav。把 World90.sav 作为当前工作文件，其数据如图 2-41 所示，它记录了 1990 年对各个国家进行调查的数据。我们需要和 1960 年的情况进行比较，因此需要按照关键变量 name 进行文件合并。在合并两个文件之前，首先要对两个数据文件按照关键变量（name）进行升序排列，这通过选择【数据】→【排序个案】来完成。这一步留给读者自己完成。只有首先完成个案的排序，然后才能进行下面的操作。

在数据文件 World90.sav 中（当前工作文件），选择【数据】→【合并文件】→【添加变量】，得到如图 2-42 所示的一对一添加变量步骤 1 对话框，选中“打开的数据集（O）”框中的数据文件“world60.sav”。

name	region	area	pop1990	lifeex87	urpop90	indeat90	educ90	inf8087
Afghanistan	Asia	65209	17	42	21.7	172	.	.
Albania	Europe	2740	3	72	35.3	39	.	.
Algeria	Africa	238174	25	63	44.7	74	6.1	5.1
Angola	Africa	124670	10	45	28.3	137	3.4	.
Argentina	South America	273669	32	71	86.2	32	3.3	298.
Australia	Oceania	761793	17	76	85.5	8	5.1	7.1
Austria	Europe	8273	8	74	57.7	11	6.0	4.1
Bahrain	Asia	68	1	71	82.9	26	.	.
Bangladesh	Asia	13391	116	52	13.6	119	2.2	11.
Barbados	N & Cent. America	43	0	75	44.7	11	.	.
Belgium	Europe	3282	10	75	96.9	10	5.6	5.
Benin	Africa	11062	5	47	42.0	110	3.5	8.
Bhutan	Asia	4700	2	49	5.3	128	.	.
Bolivia	South America	108439	7	54	51.4	110	2.4	601.1
Botswana	Africa	56673	1	59	23.6	67	9.1	8.
Brazil	South America	845651	150	65	76.9	63	3.4	166.

图 2-41 World90.sav 数据视图

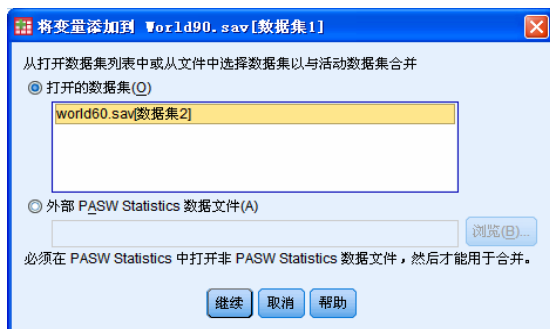


图 2-42 一对一添加变量步骤 1

单击【继续】按钮，得到如图 2-43 所示的一对一添加变量步骤 2 对话框。在该对话框中，需要设置下列事项。

- 添加变量的类型：一对一或者一对多，SPSS 默认直接把第二个文件的数据添加到当前数据文件中，而没有指定联结两个文件的关键变量。这在大部分情况下是错误的。我们需要勾选“按照排序文件中的关键变量匹配个案”。然后选择“两个文件都提供个案 (B)”，即一对一合并。另外两个选项用于指定一对多合并中的关键字表文件。
- 合并后文件中保留的变量：默认为所有非重名的变量。当前活动数据文件中的变量在前（用“*”标识）。如果某些变量不需要出现在合并后的文件中，可以把它们选入“已排除变量 (E)”框中。
- 重名变量如何处理：World90.sav 和 World60.sav 都含有变量 Name 和 Region，SPSS 默认非当前文件中的重名变量被排除，本例中为 Name 和 Region。如果合并后的文件需要这些重名的变量，需要先把这些变量重新命名，然后再选入到右边的“新的活动数据集 (N)”框中。
- 指定关键变量：把 Name 选入到“关键变量 (V)”框中。

正确完成设置后的对话框如图 2-43 所示。



图 2-43 一对一添加变量步骤 2

单击【确定】按钮，和 2.6.1 中添加个案一样，当前活动数据文件就是合并后的数据文件。如果还需要保留原来的文件，需要把当前活动数据文件另存为不同的文件名。这里另存为 World90v60.sav。

2. 一对多合并

我们用图 2-44 来说明一对多添加变量的方法。这里有两个数据文件，一个是国家文件（左边 3 栏）它含有三个变量：国家、地区和变量 1。另一个数据文件是地区数据（右边 2 栏），该文件含有两个变量：地区和变量 2。地区数据文件作为主文件，地区变量作为关键变量，得到合并变量后的数据文件（中间 3 栏）。如图 2-44 所示，国家数据文件中的前三个国家的地区变量值都是 1，因此合并后的数据文件中前三个个案都和地区数据文件的第一个个案相匹配。而最下面的 UK 和 US 都与地区文件的第三个个案匹配。

国家数据			合并后的数据			地区数据	
国家	地区	变量1	国家	变量1	变量2	地区	变量2
Canada	1	X	Canada	X	A	1	A
France	1	X	France	X	A	2	B
Spain	1	X	Spain	X	A	3	C
UK	3	X	UK	X	C		
US	3	X	US	X	C		

图 2-44 一对多添加变量

和一对一添加变量一样，一对多添加变量是通过菜单【数据】→【合并文件】→【添加变量】来完成的（如图 2-34 所示）。打开 SPSS 的两个数据文件：CustomerSurveyA.sav 和 CustomerRevenue.sav。把 CustomerSurveyA.sav 作为当前工作文件，它记录了对某软件产品 A 进行的客户满意度调查数据。CustomerRevenue.sav

是基于该产品的历史销售情况,按照工作单位性质和使用 A 产品的时间分类客户统计的年收入。这里分析的目的是比较不同收益类型客户的概要特征。这需要先按照关键变量 `orgnType` (客户工作单位的类型) 进行合并文件。

在合并两个文件之前,首先要对两个数据文件按照关键变量(`orgnType` 和 `useA`) 进行升序排列,这通过选择【数据】→【排序个案】来完成,如图 2-45 所示。

从【窗口】菜单中选择 `CustomerSurveyA.sav` 作为当前工作文件,然后选择【数据】→【合并文件】→【添加变量】,得到如图 2-46 所示的对话框,选中“打开数据集(O)”框中的文件。



图 2-45 排序个案

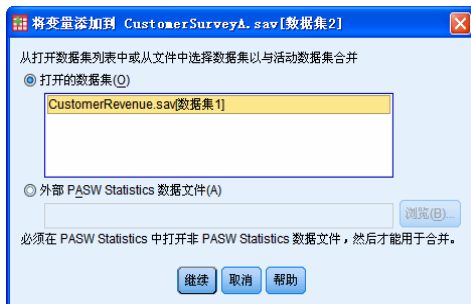


图 2-46 添加变量

单击【继续】按钮,得到如图 2-47 所示的添加变量对话框,它和图 2-43 类似,先勾选“按照排序文件中的关键变量匹配个案”前的框。不同的是,这里选择下面的第二个选项:“非活动数据集为基于关键字的表(K)”,即用 `CustomerRevenue.sav` 作为主文件(或查表文件)。其他选项和图 2-47 完全一样。

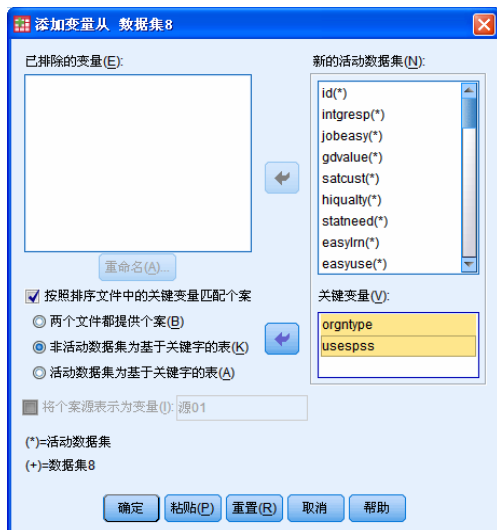


图 2-47 添加变量

(1) 把需要分析的某个组别过滤出来, 分析完该组别后关闭过滤器, 然后重复以上过程来选择和分析另一个组别, 直到分析完所有组别。

(2) 和 (1) 一样, 用【数据】→【选择个案】来分别选出各个组, 但是不进行分析, 而是把各个组别的个案都另存为数据集, 然后再分别进行分析。当需要在不同组别中进行不同的分析时, 该方法效率较高。

- 用【数据】→【拆分文件】方式。当各个组互不相交 (即没有任何个案同时归属于两个或以上的组别), 并且需要一次性对不同的组别进行相同的分析时, 该方式简单易行。本节讨论该种数据拆分方式。

以数据文件 Chapter 2 GSS04Intro.sav 为例, 变量 `marital` 记录了婚姻状况, 现在需要对不同婚姻状况的人进行比较分析。首先要按照婚姻状况对个案进行分拆。

选择【数据】→【分割文件】, 得到对话框如图 2-48 (A) 或者图 2-48 (B) 所示。默认情况下 SPSS 选择“分析所有个案, 不创建分组 (A)”。

SPSS 分割文件程序有两个选项:

- 比较组: 该选项将拆分文件组显示在一起以用于比较。对于枢轴表, 将创建单个数据枢轴表, 且可以将每个拆分文件变量在表的维度之间移动。对于图表, 为每个拆分文件组分别创建图表, 并在“浏览器”中将图表显示在一起;
- 按组组织输出: 该选项为每个拆分文件组分别显示每个过程中的所有结果。

为了便于读者理解, 我们先应用第一种方式来拆分文件, 如图 2-50 (A) 所示的分割文件对话框, 然后进行描述性统计分析, 得到如图 2-52 所示结果; 之后应用第二种方式来重新拆分文件, 如图 2-50 (B) 所示的分割文件对话框, 然后进行描述性统计分析, 得到如图 2-53 所示的结果。

按照图 2-50 (A) 和图 2-50 (B) 来分割文件之后, 运行描述性统计分析, 即【分析】→【描述性统计】→【描述】, 选择 `INCOME` 变量和 `INCOME_ACTUAL` 变量, 如图 2-51 所示。得到的分析结果分别如图 2-52 和图 2-53 所示, 前者把所有组的结果放在同一张表中, 而后者则将不同的组的分析结果输出到不同的表中。

由图 2-52 的“比较组”分隔后分析结果, 我们还可以分别列出“按组组织输出”分割后的分析结果, 如图 2-53 所示。



图 2-50 (A) 分割文件对话框——比较组



图 2-50 (B) 分割文件对话框——按组组织输出

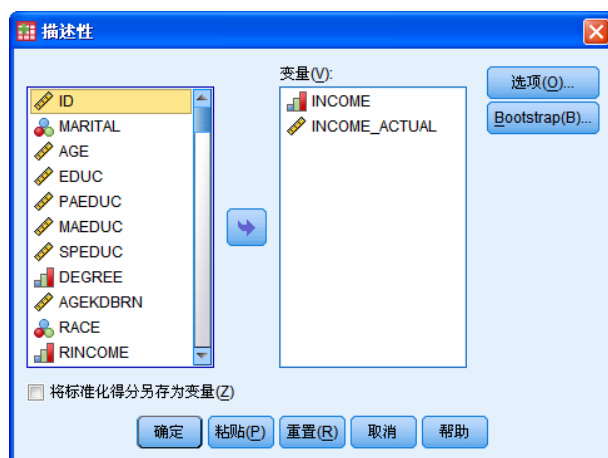


图 2-51 描述性统计分析

描述统计量						
MARITAL STATUS		N	极小值	极大值	均值	标准差
MARRIED	TOTAL FAMILY	1436.0	1.0	13.0	11.7	1.5
	Income in	902.0	560.9	99964.3	45367.1	30037.6
	有效的 N (列)	894.0				
WIDOWED	TOTAL FAMILY	183.0	1.0	13.0	10.4	2.5
	Income in	49.0	584.0	86429.3	32325.5	27779.2
	有效的 N (列)	47.0				
DIVORCED	TOTAL FAMILY	406.0	1.0	13.0	10.7	2.5
	Income in	281.0	681.4	99852.4	42221.3	27789.7
	有效的 N (列)	279.0				
SEPARATED	TOTAL FAMILY	92.0	1.0	13.0	9.9	3.3
	Income in	55.0	544.9	98615.0	34272.4	30052.0
	有效的 N (列)	54.0				
NEVER MARRIED	TOTAL FAMILY	575.0	1.0	13.0	10.1	3.2
	Income in	401.0	552.3	99717.9	34246.1	29949.3
	有效的 N (列)	384.0				

图 2-52 “比较组”分割后分析结果

MARITAL STATUS = MARRIED描述统计量^a

	N	极小值	极大值	均值	标准差
TOTAL FAMILY INCOME	1436	1	13	11.71	1.484
Income in Dollars	902	560.89	99964.29	45367.0842	30037.60983
有效的 N (列表状态)	894				

a. MARITAL STATUS = MARRIED

MARITAL STATUS = WIDOWED描述统计量^a

	N	极小值	极大值	均值	标准差
TOTAL FAMILY INCOME	183	1	13	10.38	2.460
Income in Dollars	49	583.99	86429.28	32325.5286	27779.21130
有效的 N (列表状态)	47				

a. MARITAL STATUS = WIDOWED

MARITAL STATUS = DIVORCED描述统计量^a

	N	极小值	极大值	均值	标准差
TOTAL FAMILY INCOME	406	1	13	10.69	2.539
Income in Dollars	281	681.37	99852.45	42221.3313	27789.66369
有效的 N (列表状态)	279				

图 2-53 “按组组织输出”分割后的分析结果

附录：如何为数据库文件建立 ODBC 数据源

如果操作系统为 Windows XP，在开始菜单中，选择【设置】→【控制面板】→【管理工具】→【数据源（ODBC）】，将出现如图 2-54 所示的“ODBC 数据源管理器”窗口。用户 DSN、系统 DSN 和文件 DSN 是建立用户数据源的类型，用户 DSN 意味着该 ODBC 数据源仅对当前登录到 Windows 的用户适用，而系统 DSN 则对该计算机的所有用户适用。这里，我们以默认的用户 DSN 为例，“用户数据源(U)”框中列出了已经建立的 ODBC 数据源。



图 2-54 建立 ODBC 数据源管理器窗口

单击【添加】按钮，出现如图 2-55 所示的对话框，用户从中选择数据源对应的驱动程序。这里以 MS Access 数据库为例，选择 Access Driver。

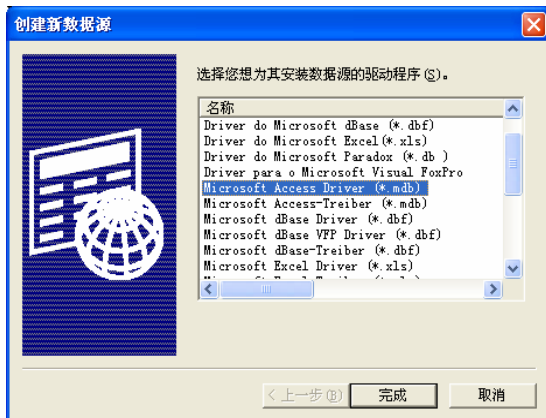


图 2-55 选择数据源驱动程序

单击【完成】按钮，出现如图 2-56 所示的对话框，要求用户输入 ODBC 数据源名称和数据源对应的数据库目录地址。



图 2-56 选择数据库

按照图 2-56 输入数据源名称“ODBC 数据源”，然后单击【选择(S)】来指定该数据源对应的数据库。这里选择 Chapter 2 GSS04.mdb，如图 2-57 所示。

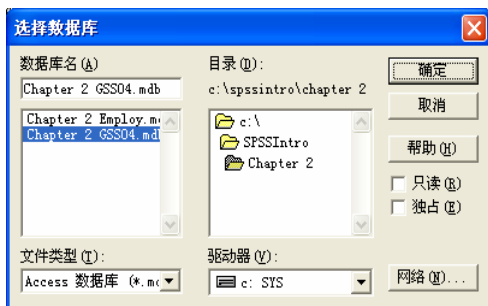


图 2-57 选择数据库

单击【确定】按钮，回到上一层对话框（如图 2-56 所示），再次单击【确定】按钮。完成 ODBC 配置后的对话框如图 2-58 所示，在用户数据源中就会出现刚刚添加的“ODBC 数据源”。



图 2-58 配置完成

单击【确定】按钮，完成 ODBC 数据源的配置，如图 2-58 所示。

2.8 小结

本章主要介绍了 SPSS 数据管理的特点。通过一个例子，从数据字典的建立到数据的输入，读者了解 SPSS 的数据编辑器：数据视图和变量视图以及在二者之间切换。SPSS 可以读入各种类型的数据文件，包括 MS Excel 数据文件、数据库文件、文本文件和其他系统生成的数据文件。本章重点介绍了如何把 Excel 数据文件、文本文件和数据库文件读入 SPSS 统计分析软件。另外，还介绍了如何把多

个数据文件根据需求合并为一个数据文件。最后介绍了如何根据分组变量对数据文件进行分割。

思考与练习

1. 一家银行的数据文件分别存储于 Excel 2003 和 Excel 2007 中，它们的文件名分别为：**BwBank.xls** 和 **BwBank.xlsx**，并且 Excel 数据表格的第一行为变量名。
 - a) 把这两个文件中的任何一个读入 **SPSS** 数据视图中。
 - b) 检查读入的数据文件，确认所有变量的数据类型为数值型，各变量的取值合理。然后，把它另存为 **SPSS** 格式的数据文件 **Newbank1.sav**。
2. 一家银行的数据文件存储于 **MS Access** 数据库中的表中，数据库名为 **BwBank.mdb**。
 - a) 把该数据库文件中的表读入 **SPSS** 数据视图中。
 - b) 检查读入的数据文件，确认所有变量的数据类型为数值型，各变量的取值合理。然后，把它另存为 **SPSS** 格式的数据文件 **Newbank2.sav**。
3. 一家银行的数据文件以分隔符分隔的文本格式保存，文件名为 **BwBank.txt**。该文本文件的第一行为变量名。
 - a) 把该文本文件读入 **SPSS** 数据视图中。
 - b) 检查读入的数据文件，确认所有变量的数据类型为数值型，各变量的取值合理。然后，把它另存为 **SPSS** 格式的数据文件 **Newbank3.sav**。
4. 有关 **SPSS** 数据字典的说法，正确的是：
 - A) **SPSS** 数据集的数据字典可以复制到其他数据集中；
 - B) **SPSS** 数据集的数据字典是不能复制的；
 - C) **SPSS** 的数据字典可以通过“复制”和“粘贴”在不同数据文件中复制。
5. **SPSS** 中可以通过多种方式查看数据字典，下列正确的是：
 - A) 通过数据编辑器的数据视图
 - B) 通过数据编辑器的变量图
 - C) 通过选择【文件】→【显示数据文件信息】
 - D) 通过选择【实用程序】→【变量】
6. 下列可以作为 **SPSS** 变量名的是：
 - A) **Prents12**

- B) 1Name
- C) NOT TRUE
- D) @result

7. SPSS 中可以设置工作目录，具体设置可以按照以下菜单：

- A) 【选项】→【设置】
- B) 【编辑】→【选项】→【设置】
- C) 【编辑】→【选项】→【文件位置】
- D) 【文件】→【选项】→【设置】

8. 当合并 Student_Infor.sav（参见表 2-1）和 Student_Scores.sav（参见表 2-2）

两个数据文件为一个数据集 Student_Records.sav 时，是增加记录还是增加变量？

- A) 增加记录
- B) 增加变量
- C) 都不是
- D) 都正确

表 2-1: Student_Infor.sav

学生 ID	性别	年龄	班级
1	Female	14	A
2	Male	15	A
3	Male	15	A
4	Female	16	B
5	Female	15	B
6	Male	15	B

表 2-2: Student_Scores.sav

学生 ID	科目	成绩
1	语文	89
2	语文	67
3	语文	78
4	语文	69
5	语文	79
1	数学	79
2	数学	84
3	数学	83
4	数学	85
5	数学	69

9. 对上题的文件合并中，哪个变量是关键变量：

- A) 学生 ID
- B) 性别
- C) 年龄和班级
- D) 科目
- E) 成绩

10. 在合并两个 SPSS 文件时, 正确的说法为:

- A) 如果是添加变量, SPSS 可以显示变量是来源于那个数据文件
- B) 如果是添加个案, SPSS 可以显示变量是来源于那个数据文件
- C) 合并两个 SPSS 文件后, 将无法辨别个案来自于哪一个文件
- D) 以上都不正确

参考文献

1. SPSS 中国公司.SPSS 初中级培训讲义.内部资料。
- 2.宋志刚等.SPSS16 实用教程.北京: 人民邮电出版社, 2008。
- 3.何丽娟等译.SPSS 统计应用与解析.北京: 电子工业出版社, 2009。

本章学习目标：

- 掌握 SPSS 数据预处理的可视离散化方法；
- 了解 SPSS 缺失值的填补方法；
- 掌握 SPSS 的数据校验方法；
- 如何标识重复个案；
- 如何标识异常个案；
- 学习如何从数据集中选择符合条件的个案。

随着计算机系统能力的提高，对信息的需要成比例增长，导致收集的数据越来越多。随之而来的问题是出现更多的个案、更多的变量以及更多的数据输入错误。这些错误会损害作为数据仓储最终目标的预测模型的预测能力，因此必须使数据保持“干净”。不过，数据仓储中数据量的增长已经大大超出了手动验证个案的能力，因而实现自动化的数据验证过程变得十分关键。

数据预处理即当录入或读取数据后，对数据进行必要的清理（包括查错纠错、标识数据中的异常个案和无效个案、变量和数据值等）、转换、填补缺失值等，为后续统计分析应用（如均值比较、方差分析、回归分析等）打下良好基础。如果把整个统计分析过程比作大厨烧菜，那么种菜或去菜场买菜等获取食材就相当于录入或读取数据，而扔掉坏的菜叶、切菜等准备工作就相当于数据预处理，而在锅里烧菜烹饪就相当于后续具体统计分析应用（如均值比较、方差分析、相关性分析、回归分析等）。可见，数据预处理虽不产生最终的分析结果，但作为最终分析的准备，是数据分析必不可少的一环，它在完整的数据分析项目过程中的位置如图 3-1 所示。

在本章中，3.1 节讨论尺度数据（即连续型数据）转换到分类数据的可视离散化方法；3.2 节讨论 SPSS 中数据缺失值的填补方法；3.3 节讨论 SPSS 中数据校验的方法；3.4 节学习如何标识重复个案和异常个案；3.5 节学习如何从数据集中选择满足

条件的个案。



图 3-1 统计分析项目过程图

3.1 可视离散化

可视离散化（可视化分段）（Visual Binning）用于为定量变量（或尺度变量）创建分类变量（或定性变量），从而实现连续变量的离散化。在统计分析中，有时候需要了解总体的大致分布状况，而不需要了解属性的具体信息。例如，调查居民的收入水平，实际得到的是以“元”计数的具体收入值。有时候用户最关心的是处于贫困线以下（假设年收入¥2 000 以下为贫困）的居民、中等收入（年收入为¥2 000-¥30 000）的居民和高收入（年收入高于¥30 000）的居民各占多大比例。这时候，可以对定量变量年收入进行“可视离散化”，创建一个包括处于贫困线以下、中等收入和高收入三个类别的新分类变量或定性变量。再比如，我们收集了居民具体的年龄数值，但我们关心的是处于各个年龄段的人群的比例。此时，可以对定量变量年龄进行“可视离散化”，创建一个包括青年、中年、老年三个类别的新分类变量。

打开数据文件 1991 U.S. General Social Survey.sav，如图 3-2 所示，该数据文件为 1991 年美国普遍社会调查数据。

在原始数据文件中，为了解各个年龄段人群的分布情况，需要对年龄变量进行可视化分段。SPSS 的可视化分段提供两类分段的方法：直接输入分割点和根据条件自动生成分割点。其中，根据条件自动生成分割点提供了三种自动生成分割点的方法：等宽间隔、基于已扫描个案的等百分位和基于已扫描个案的均值和标准差。



1991 U.S. General Social Survey.sav [数据集2] - PASW Statistics 数据编辑器

文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 应用程序(P) 图形(G) 实用程序(U) 窗口(W) 帮助

1248: 性别 2 可见: 43 变量的 43

	性别	种族	地区	幸福	生活	兄弟姐妹	子女	年龄	教育	父亲教育	母亲教育
1239	2	1	1.00	2	2	1	5	63	12	9	98
1240	2	1	1.00	1	0	10	4	35	12	8	12
1241	1	1	1.00	3	3	11	3	77	8	8	8
1242	1	1	1.00	1	0	2	0	21	12	16	13
1243	2	1	1.00	2	2	1	3	45	14	12	12
1244	2	1	1.00	2	1	1	0	22	15	14	12
1245	2	1	1.00	2	0	4	2	34	16	97	13
1246	2	1	1.00	1	2	5	2	66	14	19	16
1247	2	1	1.00	1	0	2	2	26	14	16	12
1248	2	1	1.00	3	3	9	2	32	12	12	12
1249	2	1	1.00	2	0	2	4	39	13	8	11
1250	1	1	1.00	2	2	1	1	74	12	8	8
1251	2	1	1.00	2	0	2	7	61	12	8	14
1252	2	1	1.00	2	2	1	2	35	13	8	14
1253	1	1	1.00	1	1	2	2	26	14	97	8
1254	2	1	1.00	1	0	5	7	51	12	8	16

图 3-2 1991 年美国 GSS 数据视图

注意: SPSS 中文版的可视离散化程序中对于“Cutpoints”翻译为“分割点”,有的对话框中翻译为“分隔点”,例如本章的图 3-9 中为“分割点”和图 3-10 中为“分隔点”。为保持和软件的界面一致,本书沿用 SPSS 软件界面上的用语。读者把这两个作为一个来理解即可。希望 SPSS 后续的中文版本能够更正这种翻译上的错误,避免给用户造成不必要的混淆。

3.1.1 直接输入分割点

数据文件 1991 U.S. General Social Survey.sav 中,需要进行可视化分段的变量“年龄”为定量(或尺度)变量,以下使用可视离散化(Visual Binning)对该变量进行分段并产生一个新的分类变量。

选择菜单【转换(T)】→【可视离散化(I)】,可视离散化界面如图 3-3 所示。

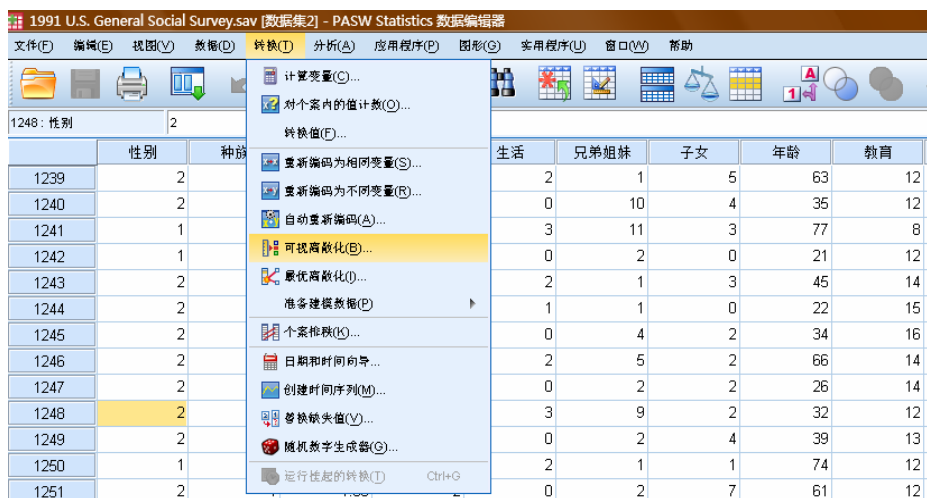


图 3-3 可视离散化界面

选择“年龄”变量进入“要离散的变量(B)”界面,如图3-4所示。

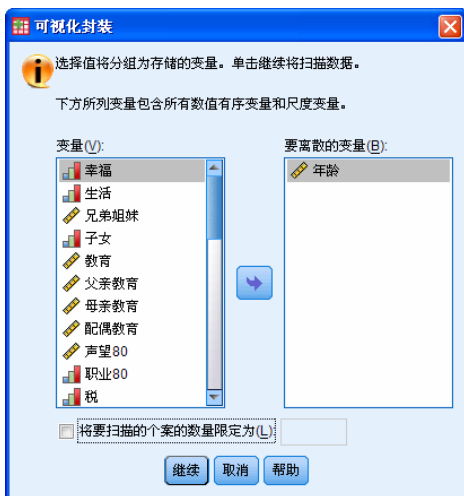


图 3-4 选择要离散的变量

单击【继续】按钮,得到如图3-5所示的可视化封装对话框。

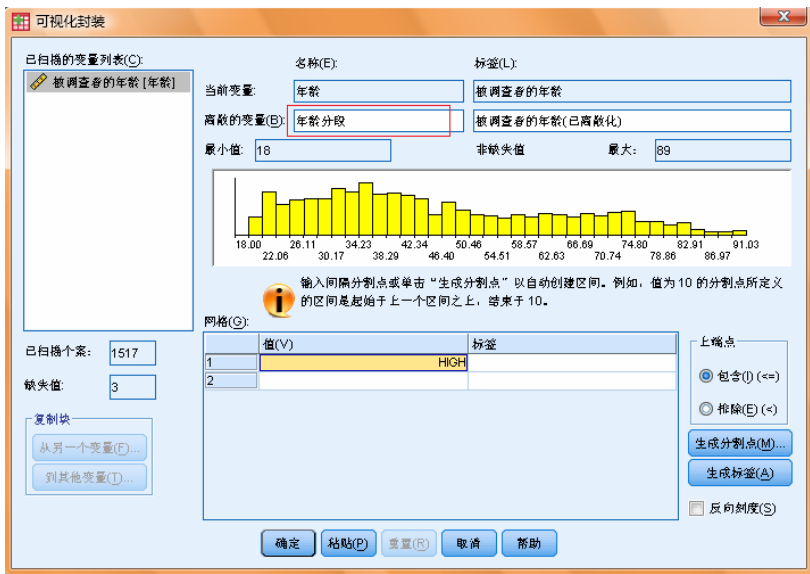


图 3-5 “可视化封装”对话框

如图3-5所示,通过中间的直方图可知年龄的分布情况,输入离散后的分类变量名称:年龄分段。假如我们把35岁及其以下的称为“青年”,35岁至52岁的称为“中年”,52岁以上的称为“老年”,即可在图3-5中“网格(G)”下面的表格中的“值(V)”所在的列输入分割点的值,在“标签”所在的列填好各年龄段所对应的标签,如图3-6所示。另外,也可以单击右下角的【生成标签(A)】按钮,自动化分段会自动生成各个分段的标签。

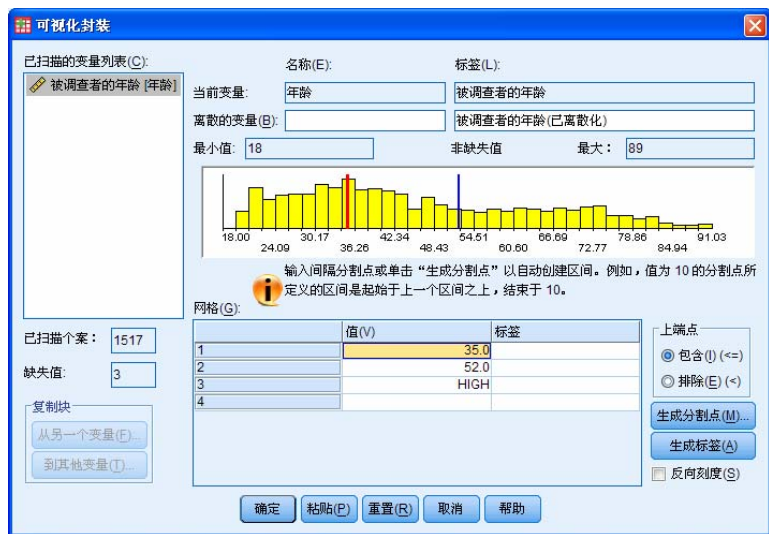


图 3-6 可视化分段的视图

输入完分割点的值之后，即可在直方图上看到相应的分割线，并由分割线及直方图可大致看出对连续变量的分段情况。本例中可把年龄分成 3 个大致含有相等百分比个案的分段。

这时如果单击【确定】按钮，将弹出如图 3-7 所示的对话框。

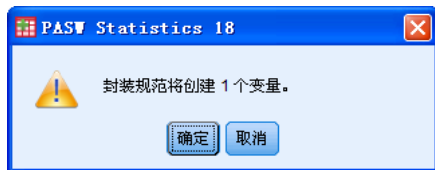


图 3-7 提醒

单击【确定】按钮，即可在数据文件中生成新变量年龄分段，如图 3-8 所示。

1: 性别	2: 工作7	工作8	工作9	问题1	问题2	问题3	问题4	年龄分段
139	AP	NAP	NAP	NAP	-	-	-	中年
140	否	否	否	否	-	-	-	老年
141	否	否	否	否	-	-	-	老年
142	AP	NAP	NAP	NAP	-	-	-	老年
143	AP	NAP	NAP	NAP	-	-	-	青年
144	否	否	否	否	健康	其他	其他	老年
145	AP	NAP	NAP	NAP	-	-	-	青年
146	否	否	否	否	-	-	-	老年
147	否	否	否	否	-	-	-	中年
148	否	否	否	否	-	-	-	青年
149	否	否	是	否	健康	金融	金融	青年
150	否	否	否	否	-	-	-	中年

图 3-8 分段后的数据视图


```

NEW FILE.

DATASET CLOSE ALL.

GET FILE = 'D:\SPSSIntro\1991 U.S. General Social Survey.sav' .

DATASET NAME myData WINDOW=FRONT.

RECODE  年龄 (MISSING=COPY)

    (LO THRU 35=1)

    (LO THRU 52=2)

    (LO THRU HI=3)

    (ELSE=SYSMIS)

INTO 年龄分段.

VARIABLE LABELS  年龄分段 '被调查者的年龄(已离散化)'.

FORMATS  年龄分段 (F5.0).

VALUE LABELS  年龄分段 1  '青年'  2  '中年'  3  '老年'  98  'DK'  99  'NA'.

MISSING VALUES  年龄分段  (0 , 98 , 99).

VARIABLE LEVEL  年龄分段 (ORDINAL).

EXECUTE.

```

可视化封装

已扫描的变量列表(C): 被调查者的年龄(年龄)

名称(E): 年龄 标签(L): 被调查者的年龄

当前变量: 年龄 离散变量: 年龄分段 被调查者的年龄(已离散化)

最小值: 18 非缺失值 最大: 89

输入间隔分割点或单击“生成分割点”以自动创建区间。例如，值为10的分割点所定义的区间是起始于上一个区间之上，结束于10。

网络(S):

值(V)	标签
1	31.0 青年
2	70.2 中年
3	HIGH 老年
4	

已扫描个案: 1517

缺失值: 3

复制块

从另一个变量(F)...

到其他变量(T)...

上端点

☒ 包含(I) (=)

☐ 排除(E) (<)

生成分割点(M)...

生成标签(A)

☐ 反向刻度(S)

确定 粘贴(P) 置入(R) 取消 帮助

3.1.2 根据条件自动生成分割点

74

在图 3-5 中，如果单击右下角的【生成分割点】按钮，则出现“生成分割点”对话框，如图 3-10 所示，在此可以输入生成分割点的条件，例如，等宽度间隔、等个案数、均值和几倍标准差。

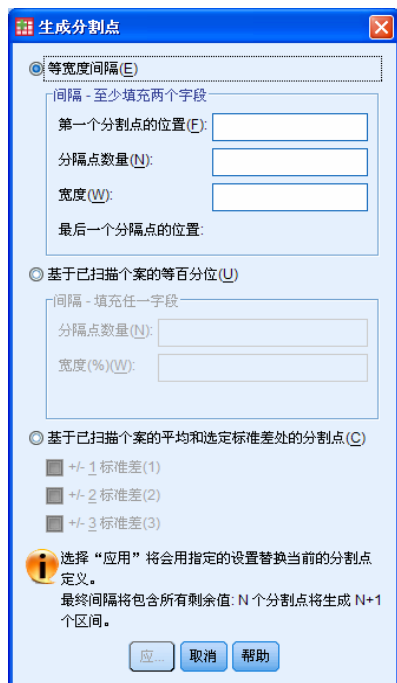


图 3-10 根据条件生成分割点

- “等宽度间隔 (E)”：输入第一个分割点位置和分割点数量，可视化分段会自动进行等间距分段。
- “基于已扫描个案的等百分位 (U)”：输入分割点数量或宽度（%），即可实现对连续变量等百分位分段。这里分隔点数量和宽度（%）只要输入一个即可，另外一个会根据输入的值自动生成。例如，如果想把年龄分为 3 段，则在“分隔点数量 (N)：”后的文本框中输入“2”，“宽度 (%) (W)”后的文本框中会自动生成 33.33%，即每个分段大约含有 33.33% 的个案。或者在宽度（%）中输入 33.33，则分隔点数量自动变为 2，如图 3-11 所示。
- “基于已扫描个案的平均和选定标准差处的分割点 (C)”：可实现根据均值和选定标准差进行分段。如果选择“+/-1 标准差”，则取均值减去 1 倍标准差、均值、均值加 1 倍标准差三个位置作为分割点，如图 3-12 所示。

这里选择如图 3-11 所示的设置，单击【应…】按钮，返回上级对话框，即图 3-6 所示的可视化封装对话窗口，在“离散的变量 (B)”后文本框内输入离散化后的变量名称“年龄段”，然后单击右下角的【生成标签 (A)】按钮，自动化分段

便自动为各分段生成相应标签：“ ≤ 35 ”、“36-52”、“53+”，分段变量预览的情况如图 3-13 所示。

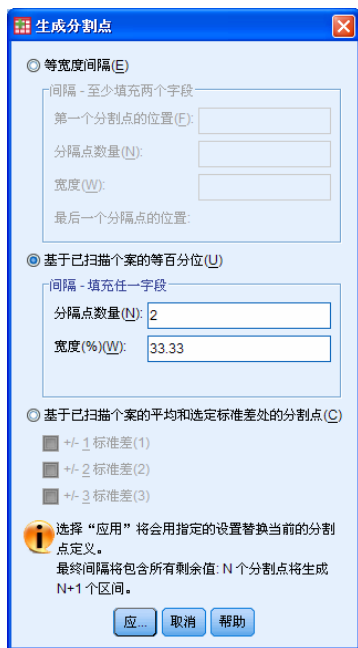


图 3-11 等百分位生成分割点

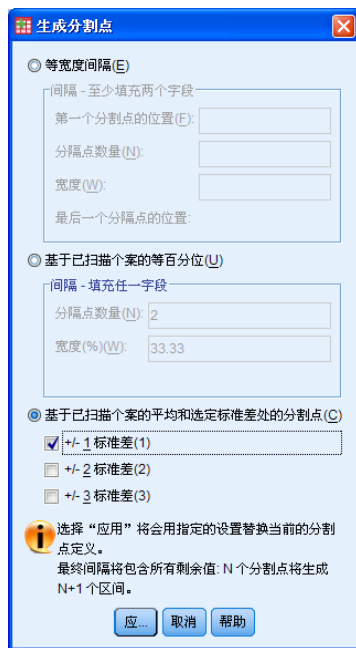


图 3-12 基于均值和标准差生成分割点

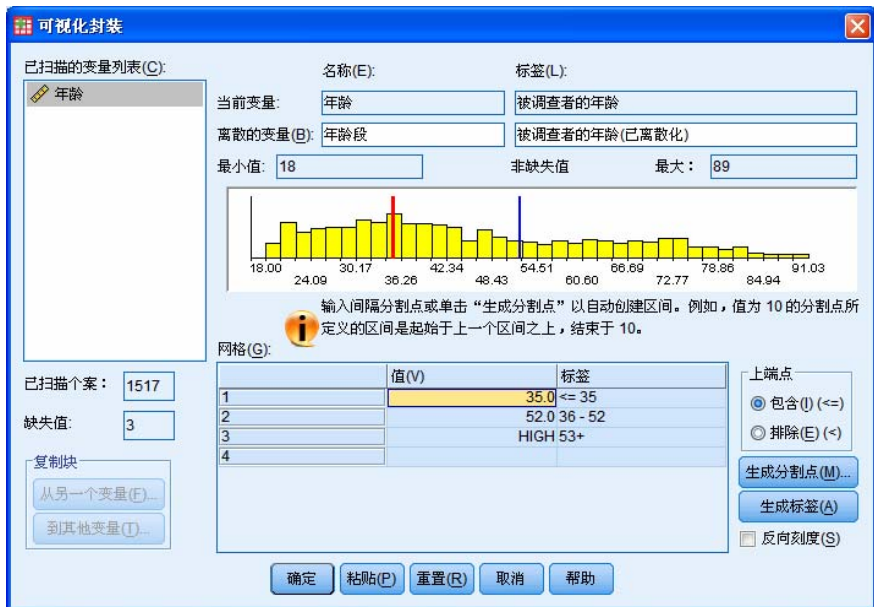


图 3-13 分段变量预览

设置完成之后，最后单击【确定】按钮，将出现如图 3-14 所示的提醒对话框。它提示用户一个名为“年龄段”的新变量将会在数据视图中生成。

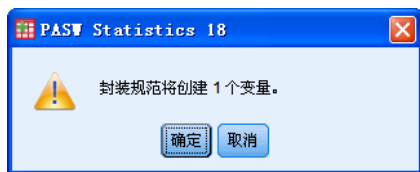


图 3-14 提醒生成新变量

单击【确定】按钮，即可完成该变量的自动化分段。

以上过程可以通过下列语法命令完成。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'D:\SPSSIntro\1991 U.S. General Social Survey.sav' .
DATASET NAME myData WINDOW=FRONT.
* 可视化封装.
* 年龄.
RECODE 年龄 (MISSING=COPY) (LO THRU 35=1) (LO THRU 52=2) (LO THRU HI=3)
(ELSE=SYSMIS) INTO 年龄段.
VARIABLE LABELS 年龄段 '被调查者的年龄(已离散化)'.
FORMATS 年龄段 (F5.0).
VALUE LABELS 年龄段 1 '<= 35' 2 '36 - 52' 3 '53+' 98 'DK' 99 'NA'.
MISSING VALUES 年龄段 (0 , 98 , 99).
VARIABLE LEVEL 年龄段 (ORDINAL).
EXECUTE.
```

完成自动化分段后，数据视图如图 3-15 所示。最右边的一列“年龄段”变量即为分段后新生成的变量。

工作7	工作8	工作9	问题1	问题2	问题3	问题4	年龄段
否	否	否	金融	金融	.	.	<= 35
否	否	否	53+
NAP	NAP	NAP	<= 35
否	否	否	53+
NAP	NAP	NAP	<= 35
否	否	否	53+
否	否	否	<= 35
NAP	NAP	NAP	36 - 52
否	否	否	36 - 52
NAP	NAP	NAP	36 - 52
否	是	否	<= 35
否	否	否	<= 35
否	否	是	<= 35
否	否	否	金融	金融	.	.	36 - 52
否	否	否	53+
NAP	NAP	NAP	36 - 52
否	否	否	36 - 52
否	否	否	36 - 52
否	否	否	53+
否	否	否	金融	.	.	.	53+

图 3-15 新分段变量

注意：在应用离散化分段时，建议仔细分析数据的直方图，了解数据的分布特点，并结合数据的具体含义和分析主题，采用相应的生成分割点的方法。

3.2 缺失值

统计分析工作者在实务中经常会碰到数据缺失的问题。一般说来，数据缺失主要由以下几种原因造成：

- 在数据收集阶段，收集者没有收集到相应数据；
- 应答者拒绝回答该问题，比如该问题涉及个人隐私；
- 该问题对该应答者不适用，比如该问题是针对女性的，而应答者为男性。

缺失数据对于分析者来说正如癌症对于医生，含有缺失数据的数据分析是不可靠的。因此对缺失数据的处理，首先应想办法重新回到数据收集阶段尽量收集到该数据；如果实在收集不到该数据，再考虑怎么处理缺失数据，如果缺失数据不影响到具体的统计分析，则不对缺失数据做任何处理（即缺失数据还是作为缺失数据处理），如果缺失数据影响到了具体的统计分析，则必须考虑采取适当方法来填补缺失数据。SPSS 统计分析软件的基本模块提供了下列填补缺失数据方法：

- 序列均值；
- 临近点均值；
- 临近点的中位数；
- 线性插值法；
- 点处的线性趋势。

打开数据文件 Cars.sav，该数据文件为不同汽车属性的数据，数据文件的第一列 mpg（每加仑汽油行驶的英里数）属性出现了较多的缺失值，如图 3-16 所示。

	mpg	引擎	马力	重量	加速	年份	原产地	汽缸
4	16	304	150	3433	12	70 年	美国	8 缸
5	17	302	140	3449	11	70 年	美国	8 缸
6	15	429	198	4341	10	70 年	美国	8 缸
7	14	454	220	4354	9	70 年	美国	8 缸
8	14	440	215	4312	9	70 年	美国	8 缸
9	14	455	225	4425	10	70 年	美国	8 缸
10	15	390	190	3650	9	70 年	美国	8 缸
11	.	133	115	3090	18	70 年	欧洲	4 缸
12	.	350	165	4142	12	70 年	美国	8 缸
13	.	351	153	4034	11	70 年	美国	8 缸
14	.	383	175	4166	11	70 年	美国	8 缸
15	.	360	175	3650	11	70 年	美国	8 缸
16	15	383	170	3563	10	70 年	美国	8 缸
17	14	340	160	3609	8	70 年	美国	8 缸
18	.	302	140	3353	8	70 年	美国	8 缸
19	15	400	150	3761	10	70 年	美国	8 缸
20	14	455	225	3086	10	70 年	美国	8 缸
21	24	113	95	2372	15	70 年	日本	4 缸

图 3-16 缺失字段视图

SPSS 基本模块提供的填补方法可以在菜单【转换(T)】→【替换缺失值(V)】中访问,如图 3-17 所示。

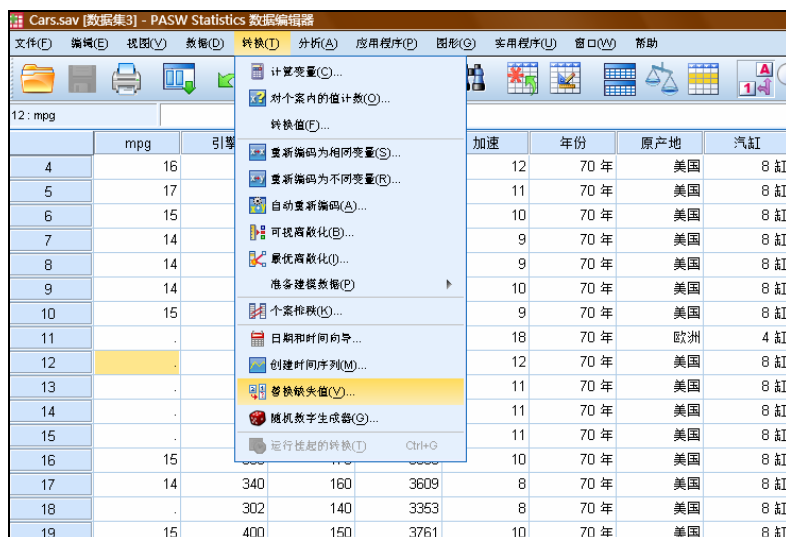


图 3-17 替换缺失值

选择菜单【转换(T)】→【替换缺失值(V)】,得到如图 3-18 所示的“替换缺失值”对话框。缺失值被替换后的变量在 SPSS 中将以一个新的变量来表示。双击需要填补缺失值的变量,该变量将作为所选定填补方法函数的参数,新变量名将赋给填补后的变量。在“方法(M)”框中列出了 SPSS 填补缺失值的五种方法。



图 3-18 选择替换缺失值的方法

在图 3-18 所示的对话框中,把变量 mpg 选到“新变量(N)”框中,填充后的变量名称默认为 mpg_1,填充方法采用默认的序列均值方法,如图 3-19 所示。设置完成后,单击【确定】按钮。

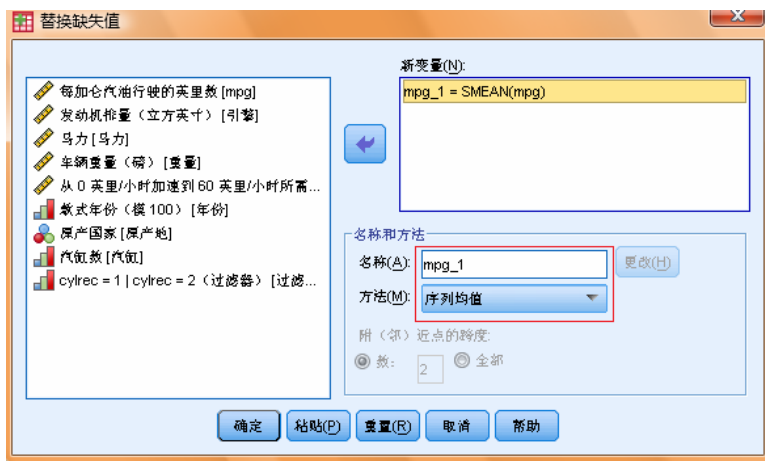


图 3-19 填补缺失值生成新变量

数据视图窗口中将出现新生成的变量 mpg_1，所有缺失值都以序列均值 23.5 填充，如图 3-20 所示。

以上操作可以通过下列语法命令完成：

```
NEW FILE.
DATASET CLOSE ALL.
GET
  FILE='D:\SPSSIntro\Cars.sav' .
DATASET NAME myData WINDOW=FRONT.
DATASET ACTIVATE myData.
RMV /mpg_1=SMEAN(mpg) .
EXECUTE.
```

*Cars.sav [数据集3] - PASW Statistics 数据编辑器											
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 应用程序(P) 图形(G) 实用程序(U) 窗口(W) 帮助											
18: mpg											
	mpg	引擎	马力	重量	加速	年份	原产地	汽缸	过滤器_\$	mpg_1	
10	15	390	190	3650	9	70 年	美国	8 缸	未选	15.0	
11		133	115	3090	18	70 年	欧洲	4 缸	已选	23.5	
12		350	165	4142	12	70 年	美国	8 缸	未选	23.5	
13		351	153	4034	11	70 年	美国	8 缸	未选	23.5	
14		383	175	4166	11	70 年	美国	8 缸	未选	23.5	
15		360	175	3850	11	70 年	美国	8 缸	未选	23.5	
16	15	383	170	3563	10	70 年	美国	8 缸	未选	15.0	
17	14	340	160	3609	8	70 年	美国	8 缸	未选	14.0	
18		302	140	3353	8	70 年	美国	8 缸	未选	23.5	
19	15	400	150	3761	10	70 年	美国	8 缸	未选	15.0	
20	14	455	225	3086	10	70 年	美国	8 缸	未选	14.0	
21	24	113	95	2372	15	70 年	日本	4 缸	已选	24.0	
22	22	198	95	2833	16	70 年	美国	6 缸	已选	22.0	
23	18	199	97	2774	16	70 年	美国	6 缸	已选	18.0	
24	21	200	85	2587	16	70 年	美国	6 缸	已选	21.0	

图 3-20 填补后的数据视图

单击 (【检索最近使用的对话框】的快捷方式) 按钮，回到刚才的“替换缺失值”窗口，即图 3-19。重复选择 mpg 进入“新变量(N)”窗口，每次采用不

同的缺失值填补方法：mpg_2 采用 2 个临近点的均值填补；mpg_3 采用 4 个临近点的中位数；mpg_4 采用线性插值法填补；mpg_5 采用点处的线性趋势法填补。选择不同填补方法后需要单击【更改 (H)】按钮，否则选定的填补方法不能生效。设置完成之后的对话框，如图 3-21 所示。



图 3-21 多种替换缺失值方法

单击【确定】按钮，回到数据视图。当前活动数据集中生成了 5 个新的经过填补的 mpg 变量：mpg_1、mpg_2、mpg_3、mpg_4、mpg_5。

以上操作可以通过下列程序来完成：

```
NEW FILE.
DATASET CLOSE ALL.
GET
  FILE='D:\SPSSIntro\Cars.sav' .
DATASET NAME myData WINDOW=FRONT.
DATASET ACTIVATE myData.
RMV /mpg_1=SMEAN(mpg)
    /mpg_2=MEAN(mpg 2)
    /mpg_3=MEDIAN(mpg 4)
    /mpg_4=LINT(mpg)
    /mpg_5=TREND(mpg) .
EXECUTE.
```

回到变量视图，把 mpg 和新生成的 5 个变量拖拽到一起进行比较，数据视图如图 3-22 所示。

	原产地	汽缸	过滤器_\$	mpg	mpg_1	mpg_2	mpg_3	mpg_4	mpg_5
7	年	美国	8 缸	未选	14	14.0	14.0	14.0	14.0
8	年	美国	8 缸	未选	14	14.0	14.0	14.0	14.0
9	年	美国	8 缸	未选	14	14.0	14.0	14.0	14.0
10	年	美国	8 缸	未选	15	15.0	15.0	15.0	15.0
11	年	欧洲	4 缸	已选	.	23.5	14.5	14.0	15.0
12	年	美国	8 缸	未选	.	23.5	14.5	14.0	15.0
13	年	美国	8 缸	未选	.	23.5	14.5	14.0	15.0
14	年	美国	8 缸	未选	.	23.5	14.5	14.0	15.0
15	年	美国	8 缸	未选	.	23.5	14.5	14.0	15.0
16	年	美国	8 缸	未选	15	15.0	15.0	15.0	15.0
17	年	美国	8 缸	未选	14	14.0	14.0	14.0	14.0
18	年	美国	8 缸	未选	.	23.5	14.5	15.0	14.5
19	年	美国	8 缸	未选	15	15.0	15.0	15.0	15.0

图 3-22 替换缺失值后数据视图

由图 3-22 可看出，不同的缺失值填补方法填补后的结果是不一样的。

- 序列均值为取整列数据的均值。
- 临近点的均值为取该缺失值临近的几个点的均值，具体几个点由附近点的跨度来设定。
- 临近点的中位数为取该缺失值临近的几个点的中位数，具体几个点由附近点的跨度来设定。
- 线性插值法应用线性插值法填补缺失值。用该列数据缺失值前一个数据和后一个数据建立插值直线，然后用缺失点在线性插值函数的函数值填充该缺失值。
- 缺失点处的线性趋势法应用缺失值所在的整个序列建立线性回归方程，然后用该回归方程在缺失点的预测值填充缺失值。

此外，更多专业的缺失值分析及填补方法，可通过 SPSS Statistics 18.0 的缺失值分析（Missing Value Analysis）模块实现，有兴趣的读者可查阅缺失值分析模块的相关资料。

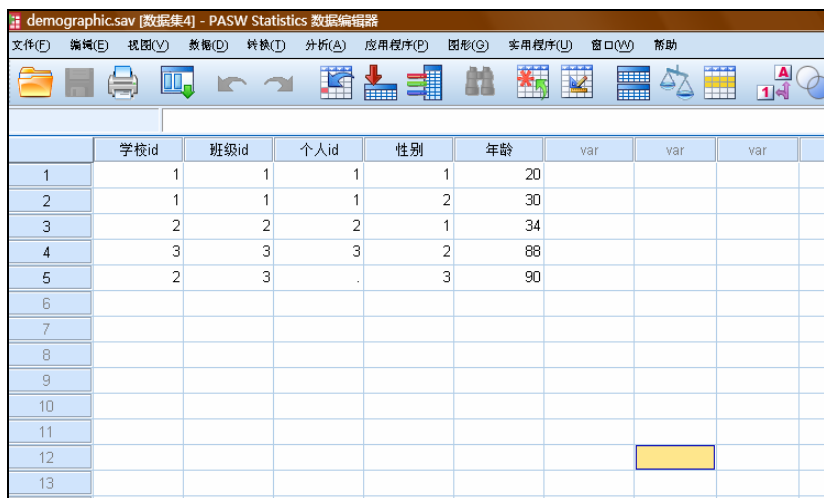
注意：如果分析中没有用到含缺失值的变量，可以不用关心缺失值问题。在 SPSS 相关的分析过程中，选择“按对排除个案（P）”，这时如果没有用到含缺失值的变量，缺失值对分析没有影响；如果选择“按列表排除个案（L）”，含有缺失值的个案将不会用于分析，可能会造成信息损失。

3.3 数据校验

一般说来，在做统计分析之前都会先做数据校验，即找出错误数据并查找错误出现的原因。如果数据没有收集到则尽量想办法补全；如果是录入错误则重新录入；

如果数据确实错误，则可将这些数据设置成缺失值（即丢弃这些数据不进入分析）。我们称查找错误数据或者不一致数据的过程为数据校验。如何使数据校验过程流程化、标准化、并且可以重复进行？SPSS Statistics 18.0 的数据准备（Data Preparation）模块为我们提供了方便的数据校验功能。

打开数据文件 demographic.sav，该数据集含有 5 条记录，每条记录包括学校 id、班级 id、个人 id 三个标识变量，还包括性别（其中“1”代表男性，“2”代表女性）和年龄两个变量，数据视图如图 3-23 所示。



	学校id	班级id	个人id	性别	年龄	var	var	var
1	1	1	1	1	20			
2	1	1	1	2	30			
3	2	2	2	1	34			
4	3	3	3	2	88			
5	2	3	.	3	90			
6								
7								
8								
9								
10								
11								
12								
13								

图 3-23 数据视图

为示例数据校验功能，这里定义 18~70 岁为有效值，而该范围之外的取值作为无效值处理。从这 5 个个案可以看出，个案 1 和个案 2 的三个标识变量重复，即学校 id、班级 id、个人 id 完全一样。记录 4 的年龄为无效值（大于 70 岁）。记录 5 的个人 id 缺失，且该记录的性别和年龄均为无效值。通过 SPSS Statistics 18 的数据准备（Data Preparation）功能，可以实现数据校验，找到无效或者错误的记录。

数据验证功能可以通过菜单路径【数据（D）】→【验证（L）】→【验证数据（V）】访问，如图 3-24 所示。

选择【数据（D）】→【验证（L）】→【验证数据（V）】，得到如图 3-25 所示的“验证数据”对话框，默认显示的标签为“变量”标签。对“分析变量（A）”和“个案标识变量（C）”进行如图 3-25 所示的设置。

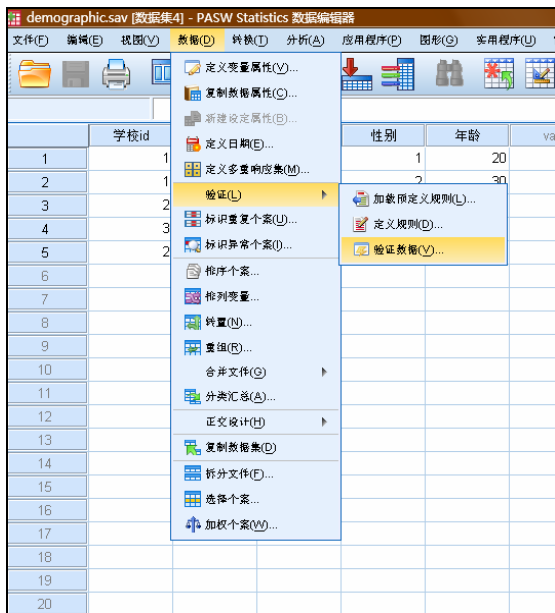


图 3-24 数据验证



图 3-25 “验证数据”对话框

其中“个案标识变量 (C)”为辨识个案（或记录）的变量，相当于数据库中的主键，它可以唯一的标示一个个案，不同个案的个案标识值必须完全不同；“分析变量 (A)”为定义验证规则并对其进行校验的变量。

“基本检查”标签中的设置保持不变，单击“单变量规则”标签，出现验证数据对话框，如图 3-26 所示。

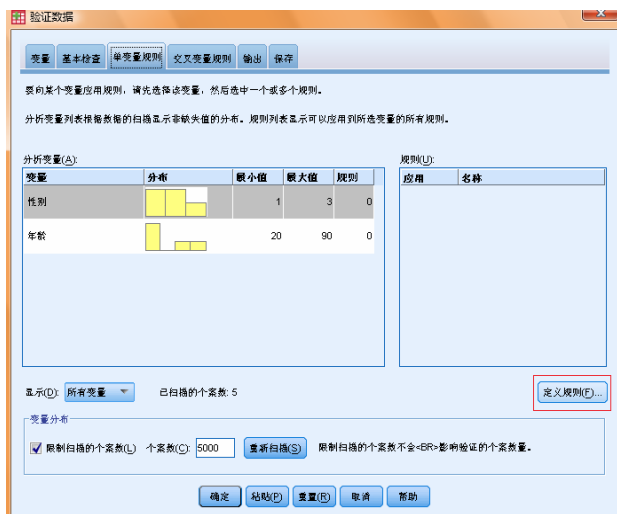


图 3-26 单变量规则

单击图 3-26 右下角的【定义规则 (F)】按钮，得到如图 3-27 所示的定义验证规则对话框。在“规则定义”框中定义两个分析变量的规则：

- 定义性别变量的规则

在“规则定义”部分：

“名称 (M)”后的文本框中输入规则名称：性别规则，类型为数字。

“有效值 (V)”：在下拉列表中选择“在列表中”，在“值 (L)”表格中输入“1”和“2”。即有效的性别值只能是“1”（男性）或者“2”（女性），如图 3-27 所示。



图 3-27 定义验证规则——单变量规则

- 定义年龄变量的规则

在图 3-27 的定义验证规则中，单击左下角【新建 (N)】按钮，类似地定义年龄规则。

在“规则定义”部分：

名称 (M)：“年龄规则”，“类型 (T)”：数字

在“有效值 (V)”部分：

在下拉列表中选择：“在范围内”，

最小值 (I)：18；最大值 (X)：70



图 3-28 单变量规则定义

完成如图 3-28 所示的设置后，单击【继续】按钮，返回到图 3-26 所示的单变量规则对话框。此时，新定义的两个规则出现在右端的“规则 (R)”框中，如图 3-29 所示。先选中“分析变量 (A)”框中的“性别”，并对性别变量应用刚才定义的性别规则，即在“规则 (U)”框中的“性别规则”前打钩。

然后，选中“分析变量 (A)”框中的“年龄”变量，并对年龄变量应用刚才定义的年龄规则，即在“年龄规则”前打钩，如图 3-30 所示。

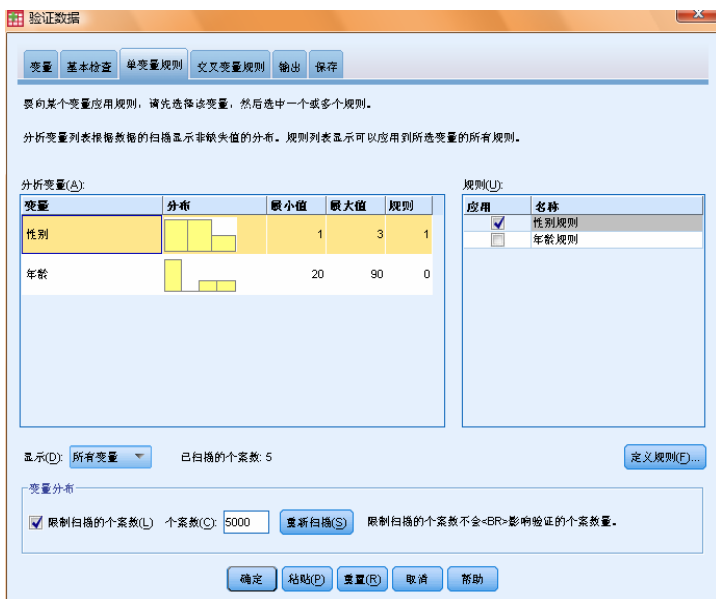


图 3-29 应用性别规则



图 3-30 应用年龄规则

“交叉变量规则”主要是针对两个或两个以上的分析变量定义规则，比如在本例中定义“性别为男性而且年龄在 18~30 岁”的个案为有效个案，这里保持“交叉变量规则”标签和“输出”标签的选项设置不变。

单击“保存”标签，“摘要变量(S)”框中的第二列“保存”用于选择是否把相应的第一列中的“描述”指标保存在数据文件中。这里选中保存所有四个指标，即勾选全部四个复选框，如图 3-31 所示。

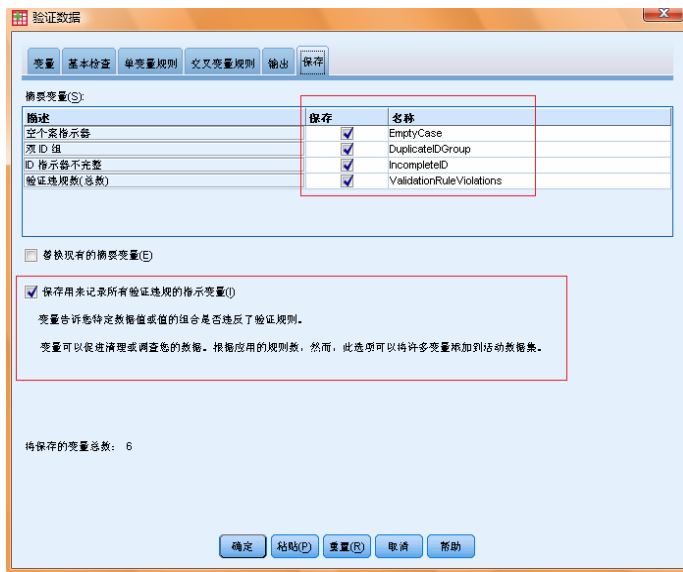


图 3-31 保存选项

单击【确定】按钮，在输出浏览器中显示下列结果。

1. 标识符检查

1) 不完全的标识符

表 3-1 显示具有不完全标识符的个案。

表 3-1 不完全的标识符

案例	标识符		
	学校 id	班级 id	个人 id
5	2	3	.

这一部分结果是检查标识变量，其中个案 5 的标识变量不完全。其标识符分别为：学校 id 值为 2，班级 id 值为 3，个人 id 为空值。因此被显示为不完全标识符个案。

2) 重复的标识符

表 3-2 显示具有重复标识符的个案。

表 3-2 重复的标识符

重复的标识符组	重复数	具有重复标识符的个案	标识符		
			学校 id	班级 id	个人 id
1	2	1, 2	1	1	1

它显示个案 1 的标识符重复了 2 次，重复的两个个案为个案 1 和个案 2。

3) 规则描述

这一部分是对单变量规则的总结，显示数据校验所应用的规则的情况。首先对性别规则和年龄规则进行描述。表 3-3 显示数据校验所应用的规则。

表 3-3 规则描述

规则	描述
年龄规则	类型: 数字 域: 范围 标记用户缺失值: 否 标记系统缺失值: 否 极小值: 18 极大值: 70 标记范围内未标记的值: 否 标记范围内的非整数値: 否 \$VD.SRule[1]: 规则
性别规则	类型: 数字 域: 列表 标记用户缺失值: 否 标记系统缺失值: 否 列表: 1, 2 \$VD.SRule[2]: 规则

显示至少违反一次的规则。

4) 变量摘要

表 3-4 为变量违反规则总结。

表 3-4 变量摘要

	规则	违规数
性别	性别	1
	总计	1
年龄	年龄	2
	总计	2

变量摘要总结数据中违反性别规则的个案数目为 1，违反年龄规则的个案数目为 2。

5) 个案报告

表 3-5 为个案违反规则情况报告。

表 3-5 个案报告

案例	确认违反规则	标识符		
	单变量 a	学校 id	班级 id	个人 id
4	年龄 (1)	3	3	3
5	年龄 (1) 性别 (1)	2	3	.

表 3-5 给出详细的个案报告，个案 4 违反了年龄规则，个案 5 违反了年龄规则和性别规则。

返回到数据视图窗口，如图 3-32 所示。

	学校id	班级id	个人id	性别	年龄	年龄规则_年龄	性别规则_性别	EmptyCase	IncompleteID	DuplicateIDGroup	ValidationRuleViolations
1	1	1	1	1	20	0	0	0	0	1	0
2	1	1	1	2	30	0	0	0	0	1	0
3	2	2	2	1	34	0	0	0	0	0	0
4	3	3	3	2	88	1	0	0	0	0	1
5	2	3	.	3	90	1	1	0	1	0	2
6											

图 3-32 个案违反规则报告

由图 3-32 可看出，新生成的六个变量提供了详细的数据校验信息，其中“1”代表无效（即违反规则），“0”代表有效（即符合规则）。

“年龄规则_年龄”变量说明个案 4 和 5 违反了年龄规则，“性别规则_性别”变量说明个案 5 违反了性别规则，EmptyCase 说明没有个案的标识变量全部为空，IncompleteID 变量说明个案 5 的标识变量不完全，DuplicateIDGroup 变量说明个案 1 和个案 2 的标识变量出现重复，ValidationRuleViolations 变量说明个案分析变量的违规数，其中个案 4 违规数为 1（违反了年龄规则），个案 5 违规数为 2（违反了年龄规则和性别规则）。

本示例仅用于说明数据校验的操作，所以相对简单，只有 5 个变量和 5 个个案，在实际工作中，如果数据量较大，个案数较多（比如几万条个案），变量数较多（比如几百个变量），那么运用 SPSS Statistics 18 的数据准备（Data Preparation）模块

就可以快速简便地实现数据的校验。因此，该模块是数据预处理中非常实用的一个模块。

3.4 标识重复个案和异常个案

3.4.1 标识重复个案

当输入大量数据时，有时候会意外地出现输入同一条记录多次；或同一条记录的某部分多次出现，即多个个案具有相同的主标识值，但它们有不同的次标识值（比如，同一个身份证号有两个不同的性别）。另外一种出现重复个案的情况是，多个个案代表同一个案，但是除这些个案的标识变量取值相同之外，其他变量的取值不同（比如，由同一个人在不同时间购买的不同产品）。

SPSS 的“标识重复个案”可以由用户对“重复个案”进行定义，并在一定程度上控制对主个案和重复个案的自动确定。这样可以找出输入数据的“意外错误”，同时也可以找出那些符合给定“重复条件”的个案。

本节仍然以 3.3 节中的数据文件为例。打开数据文件 `demographic.sav`，然后在菜单中选择【数据】→【标识重复个案】，然后把“学校 id”，“班级 id”，“个人 id”选入“定义匹配个案的依据（D）”框中，如图 3-33 所示。



图 3-33 标识重复个案

单击【确定】按钮，在输出查看器中得到如下的重复个案报告。

表 3-6 重复个案报告

所有最后一个匹配个案的指示符为主个案

		频率	百分比	有效百分比	累积百分比
有效	重复个案	1	20.0	20.0	20.0
	主个案	4	80.0	80.0	100.0
	合计	5	100.0	100.0	

从表 3-6 可知，合计部分显示数据文件一共有 5 个个案，其中有一个个案和其他个案重复，重复个案占总个案的 20%。然后，回到数据视图中，如图 3-34 所示，有一个新生成变量“最后一个基本个案”，它标识个案 1 和个案 2 是重复的，其中个案 2 被标识为主个案。

	学校id	班级id	个人id	性别	年龄	最后一个基本个案
1	1	1	1	1	20	重复个案
2	1	1	1	2	30	主个案
3	2	2	2	1	34	主个案
4	2	3	.	3	90	主个案
5	3	3	3	2	88	主个案
6						

图 3-34 重复个案报告

注意：在判定重复个案出现的原因时，建议把分析 SPSS 重复个案报告与数据的具体意义相结合，为以后的数据收集和录入提供指导。

3.4.2 标识异常个案

“标识异常个案”过程基于个案偏离聚类组中心的大小来判断异常个案。该过程一般应用于探索性数据分析步骤中，可以快速地检测到数据审核中的异常个案，它优先于任何推论性数据分析过程。此算法设计为一般“异常检测”，即异常个案的定义不被指定为任何特定应用程序。例如对医疗保健行业中异常付款模式的检测或对金融业中洗钱行为的检测，其中对异常的定义可以被很好地界定。

异常探测程序可以分为三个阶段：

- 建模阶段：根据输入的变量，进行聚类分析，找出数据集中没有明显标识出的自然组别。同时，把聚类分析最后得到的聚类模型和聚类组别的类中心保存下来。
- 打分阶段：根据建模阶段的聚类模型把个案划分到各个类，同时创建衡量每个个案偏离其所在类中心的指标变量，这里称为异常指标。异常指标最大的个案将被识别为异常个案。

- 推论阶段：根据打分阶段异常指标的大小对个案进行降序排列。一定比例的排序居前个案的异常指标值和其他变量的取值将出现在输出结果中。用户将根据这些值来判断该个案为什么被判定为异常个案。

比如，中风治疗效果分析中，数据分析人员对数据质量非常关注，因为这类模型对数据异常值十分敏感。某些被判为异常个案的观察值可能是实际有效的个案，而其值却不同于大部分其他个案的取值，因而该个案不能被用于预测分析建模。有些被判为异常个案的观察值是由于数据输入错误导致的，但是这些个案的变量取值从技术上说是“正确”的，因而不能被数据验证过程捕获。该信息收集在 `stroke_valid.sav` 中。本节将介绍如何使用“标识异常个案”过程剔除异常个案，使数据文件变得干净。

选择【数据】→【标识异常个案】打开如图 3-35 所示的“标识异常个案”的菜单，如图 3-36 所示。

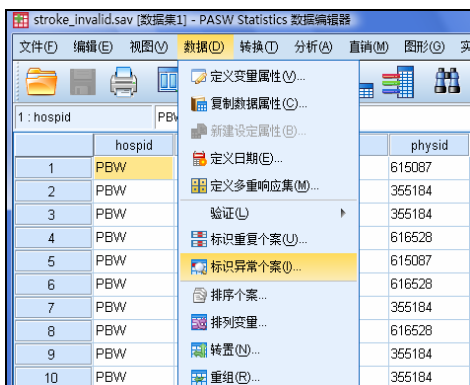


图 3-35 标识异常个案菜单

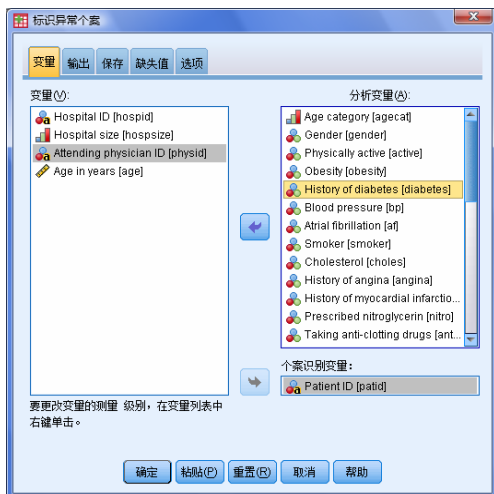


图 3-36 “标识异常个案”——变量对话框

在图 3-36 所示的“标识异常个案”——变量对话框中有 5 个标签项。

1) “变量”标签设置用于异常探测的变量, 这里将变量“Patient ID”选入对话框右下边的“个案识别变量”框中, 把“变量(V)”列表中的变量“Age category”到变量“Stroke between 3 and 6 months”选入“分析变量(A)”中。

2) 在“输出”标签中将“对等组标准值”、“异常指标”、“按分析变量列出出现的原因”、“已处理个案”都打上钩, 如图 3-37 所示。

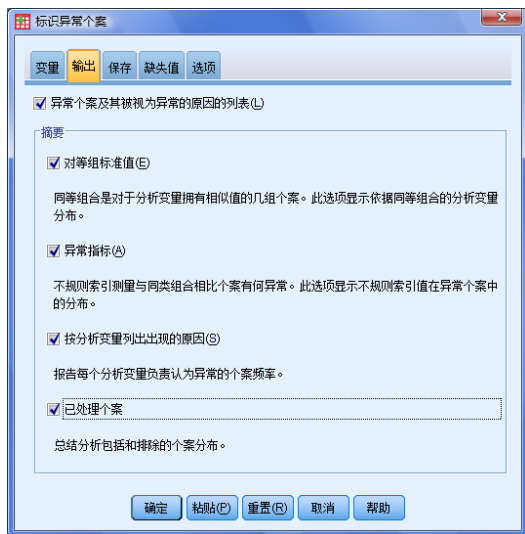


图 3-37 “标识异常个案”——输出对话框

3) 在“保存”标签中将“异常指标”、“对等组”、“原因”都打上钩, 如图 3-38 所示。



图 3-38 “标识异常个案”——保存对话框

4) 在“缺失值”标签中选择“在分析中包括缺失值(I)”，如图 3-39 所示。

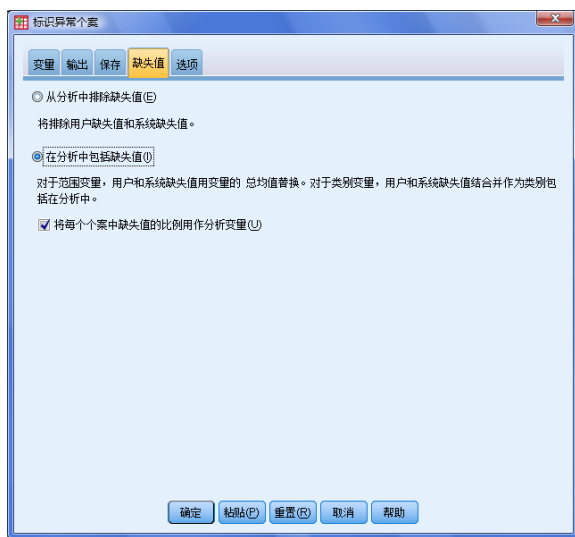


图 3-39 “标识异常个案”——缺失值对话框

该操作将把变量的缺失值用均值替换，以求最大化地利用数据。

5) 将“选项”标签中的“具有最高异常指标值的个案所占的百分比”设定为 2，将“最大的原因数量”设定为 3，如图 3-40 所示。

单击【确定】按钮。

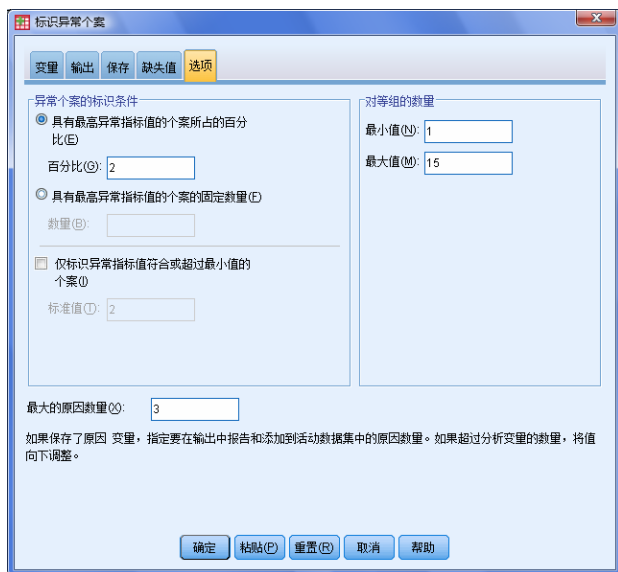


图 3-40 “标识异常个案”——选项对话框

以上操作可以通过下列的语法命令程序来实现：

```
*Identify Unusual Cases.
DETECTANOMALY /VARIABLES
    CATEGORICAL=agecat gender active obesity diabetes bp af smoker choles
    angina mi nitro anticlot tia time doa rankin0 catscan clotsolv dhosp result
    surgery rehab rankin1 rankin2 rankin3 barthell1 barthell2 barthell3 recbart1
    recbart2 recbart3 stroke1 stroke2 stroke3
    CONTINUOUS=los_rehab cost
    ID=patid
/PRINT ANOMALYLIST NORMS ANOMALYSUMMARY REASONSUMMARY CPS
/SAVE ANOMALY(AnomalyIndex) PEERID(PeerId) PEERSIZE(PeerSize)
    PEERPCTSIZE(PeerPctSize) REASONVAR(ReasonVar) REASONMEASURE(ReasonMeasure)
    REASONVALUE(ReasonValue) REASONNORM(ReasonNorm)
/HANDLEMISSING APPLY=YES CREATEMISPROVAR=YES
/CRITERIA    PCTANOMALOUSCASES=2    ANOMALYCUTPOINT=NONE    MINNUMPEERS=1
MAXNUMPEERS=15 NUMREASONS=3.
```

在输出浏览器中，得到从表 3-7 到表 3-12 所示的输出结果。

“标识异常个案”采用的聚类算法是两步聚类算法，通过两步聚类将所有个案分配到某特定的类中（即对等组）中。

表 3-7 个案处理总结（Case Processing Summary）

		N	% of Combined	% of Total
Peer ID	1	539	45.6%	45.6%
	2	644	54.4%	54.4%
Combined		1183	100.0%	100.0%
Total		1183		100.0%

在本例中，将 1183 个个案分成两类，即两个对等组。

异常个案指标列表（如表 3-8 所示）列出了所有个案中 2% 的 24 个异常个案。

表 3-8 异常指标列表（Anomaly Case Index List）

Case	Patient ID	Anomaly Index
865	3941245019	1.591
1074	3895317548	1.586
244	2711415337	1.571
148	2060299046	1.571
558	7694792176	1.559
626	4544033081	1.541
627	4544033081	1.541
628	4544033081	1.541
581	7516953949	1.536
471	6003566333	1.529
1056	5160591589	1.513

续表

Case	Patient ID	Anomaly Index
209	6376113400	1.513
981	6282443492	1.505
290	4380514785	1.497
602	7801966832	1.491
207	3307213712	1.490
163	1767580169	1.480
1094	0643047416	1.479
905	4837780819	1.479
314	6374974283	1.438
538	6390301750	1.430
503	8650135429	1.430
945	4751537050	1.425
186	7080054743	1.423

其中异常指标值（Anomaly Index）是计算每个个案与其所在类中心的距离，该值越大，说明该个案越异常，同时表 3-9 也列出了每个异常个案的个案编号和“Patient ID”，便于用户查阅异常个案。

表 3-9 个案组所属的聚类组（Anomaly Case Peer ID List）

Case	Patient ID	Peer ID	Peer Size	Peer Size Percent
865	3941245019	1	539	45.6%
1074	3895317548	1	539	45.6%
244	2711415337	1	539	45.6%
148	2060299046	1	539	45.6%
558	7694792176	1	539	45.6%
626	4544033081	1	539	45.6%
627	4544033081	1	539	45.6%
628	4544033081	1	539	45.6%
581	7516953949	1	539	45.6%
471	6003566333	1	539	45.6%
1056	5160591589	1	539	45.6%
209	6376113400	1	539	45.6%
981	6282443492	1	539	45.6%
290	4380514785	1	539	45.6%
602	7801966832	1	539	45.6%
207	3307213712	1	539	45.6%
163	1767580169	1	539	45.6%

续表

Case	Patient ID	Peer ID	Peer Size	Peer Size Percent
1094	0643047416	1	539	45.6%
905	4837780819	1	539	45.6%
314	6374974283	1	539	45.6%
538	6390301750	1	539	45.6%
503	8650135429	1	539	45.6%
945	4751537050	1	539	45.6%
186	7080054743	1	539	45.6%

表 3-10 列出了 24 个异常个案所属的聚类组，聚类组的个案数和占总体的百分比。

表 3-10 异常个案原因列表（Anomaly Case Reason List）

Reason:1

Case	Patient ID	Reason Variable	Variable Impact	Variable Value	Variable Norm
865	3941245019	Result	.110	3	1
1074	3895317548	barthel1	.136	45	90
244	2711415337	barthel1	.125	50	90
148	2060299046	Diabetes	.069	1	0
558	7694792176	barthel1	.097	60	90
626	4544033081	Rankin3	.074	(Missing Value)	0
627	4544033081	Rankin3	.074	(Missing Value)	0
628	4544033081	Rankin3	.074	(Missing Value)	0
581	7516953949	Rankin3	.075	(Missing Value)	0
471	6003566333	Rankin3	.075	(Missing Value)	0
1056	5160591589	barthel2	.084	75	100
209	6376113400	barthel3	.093	80	100
981	6282443492	barthel1	.075	65	90
290	4380514785	barthel3	.108	70	100
602	7801966832	barthel1	.112	55	90
207	3307213712	barthel3	.094	80	100
163	1767580169	barthel3	.095	80	100
1094	0643047416	barthel2	.113	70	100
905	4837780819	barthel2	.086	75	100
314	6374974283	Rankin3	.080	(Missing Value)	0
538	6390301750	barthel1	.106	60	90
503	8650135429	barthel3	.098	80	100
945	4751537050	barthel2	.118	70	100
186	7080054743	Rankin3	.080	(Missing Value)	0

异常个案原因列表列出了导致该个案为异常个案的最主要原因变量及其影响系数、该个案在变量的值以及该变量的正常值。通过追溯原因，可知该个案之所以被判定为异常个案主要是由于该个案的哪些变量值引起的。

注意：建议仔细分析异常个案的异常原因列表，结合数据的具体意义和分析的主题，判断该个案是否应该被排除在分析之外。

表 3-11 列出了定量变量在每个对等组上的均值和标准差，该均值和标准差可视为每个对等组在该变量上的聚类中心。

表 3-11 定量变量的组中心和标准差（Scale Variable Norms）

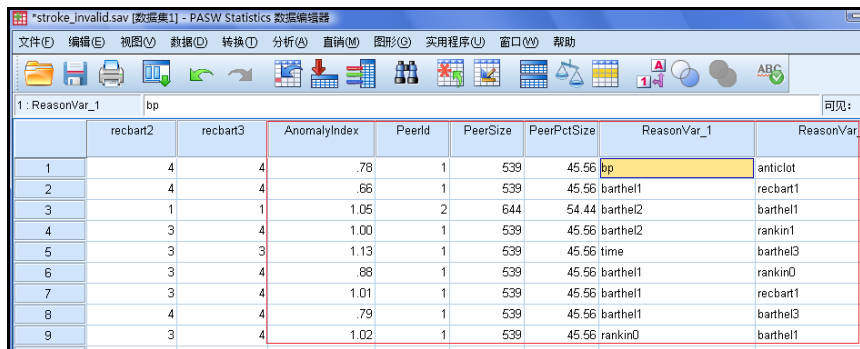
		Peer ID		Combined
		1	2	
Length of stay for rehabilitation	Mean	16.39	18.71	17.65
	Std. Deviation	12.738	12.466	12.638
Total treatment and rehabilitation costs in thousands	Mean	38.3036	50.5324	44.9607
	Std. Deviation	20.71257	30.20942	26.99713
Missing Proportion	Mean	.004	.065	.037
	Std. Deviation	.018	.093	.076

表 3-12 列出了分类变量在每个对等组上的众数（即所占百分比最高的取值），该众数可视为每个对等组在该变量上的聚类中心。

表 3-12 分类变量的聚类中心（Categorical Variable Norms）

		Peer ID		Combined
		1	2	
Age category	Most Popular Category	2	2	2
	Frequency	209	215	424
	Percent	38.8%	33.4%	35.8%
Gender	Most Popular Category	1	0	0
	Frequency	275	328	592
	Percent	51.0%	50.9%	50.0%
Physically active	Most Popular Category	1	0	0
	Frequency	285	342	596
	Percent	52.9%	53.1%	50.4%

回到在当前活动数据集中，注意到每个个案生成了 **Anomaly Index**（异常指数）、**Peer ID**（对等组 ID）、**Peer Size**（对等组个案总数）等变量，方便用户查看每个个案异常指数的详细信息，如图 3-41 所示的“标识异常个案”报告。



The screenshot shows the 'Identify Outliers' report in SPSS. The report lists 9 cases with their respective outlier indices and reasons. Case 1 is highlighted as an outlier with an index of .78 and reason 'bp'.

	recbart2	recbart3	AnomalyIndex	PeerId	PeerSize	PeerPctSize	ReasonVar_1	ReasonVar_2
1	4	4	.78	1	539	45.56	bp	anticlot
2	4	4	.66	1	539	45.56	barthel1	recbart1
3	1	1	1.05	2	644	54.44	barthel2	barthel1
4	3	4	1.00	1	539	45.56	barthel2	rankin1
5	3	3	1.13	1	539	45.56	time	barthel3
6	3	4	.88	1	539	45.56	barthel1	rankin0
7	3	4	1.01	1	539	45.56	barthel1	recbart1
8	4	4	.79	1	539	45.56	barthel1	barthel3
9	3	4	1.02	1	539	45.56	rankin0	barthel1

图 3-41 “标识异常个案”报告

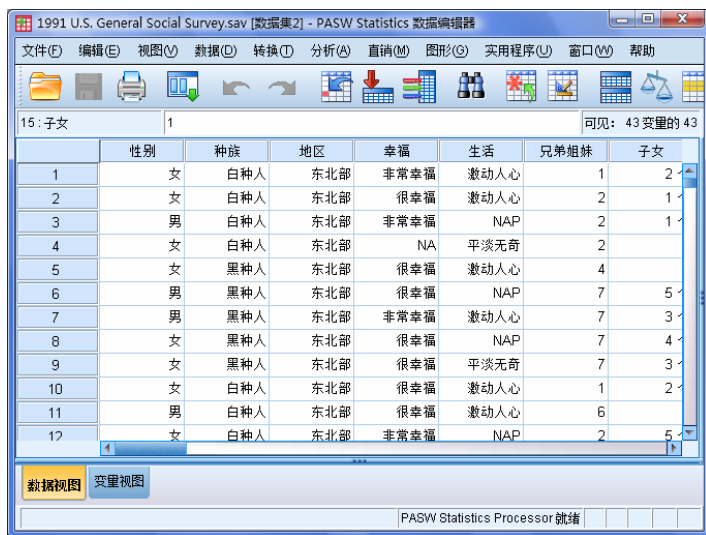
总之，通过“标识异常个案”报告，可快速挑选出异常个案。“标识异常个案”方法既可以用于建模前的数据准备，也可用于金融行业的欺诈监测（Fraud Detection），如信用卡欺诈的监测、税收欺诈的监测等。

3.5 选择个案

“选择个案”可选择数据集中特定的个案或记录，并对所选择的个案或记录进行分析。

需要特别注意的是，进行了“选择个案”的操作后，之后所有的分析所针对的个案都是基于所选择的特定个案或记录，直到取消选择个案。

打开数据文件 1991 U.S. General Social Survey.sav，该数据文件为 1991 年美国普遍社会调查的数据，数据视图如图 3-42 所示。



The screenshot shows the 'Data View' of the '1991 U.S. General Social Survey.sav' dataset. The table displays 12 cases with columns for gender, race, region, happiness, life, siblings, and children.

	性别	种族	地区	幸福	生活	兄弟姐妹	子女
1	女	白种人	东北部	非常幸福	激动人心	1	2
2	女	白种人	东北部	很幸福	激动人心	2	1
3	男	白种人	东北部	非常幸福	NAP	2	1
4	女	白种人	东北部	NA	平淡无奇	2	
5	女	黑种人	东北部	很幸福	激动人心	4	
6	男	黑种人	东北部	很幸福	NAP	7	5
7	男	黑种人	东北部	非常幸福	激动人心	7	3
8	女	黑种人	东北部	很幸福	NAP	7	4
9	女	黑种人	东北部	很幸福	平淡无奇	7	3
10	女	白种人	东北部	很幸福	激动人心	1	2
11	男	白种人	东北部	很幸福	激动人心	6	
12	女	白种人	东北部	非常幸福	NAP	2	5

图 3-42 数据视图

这里我们选择变量性别、种族和地区分别为“男性、黑种人、东北部地区”的个案。选择菜单【数据(D)】→【选择个案】，得到如图 3-43 所示的“选择个案”对话框。



图 3-43 “选择个案”对话框

单击“如果条件满足(C)”下方的【如果(I)】按钮，并在图 3-44 所示的“选择个案 If”对话框的“文本输入框”中构建选择“男性、黑种人、东北部地区”个案的表达式。



图 3-44 “选择个案——选择条件”对话框

单击【继续】按钮返回上级对话框，如图 3-43 所示。然后单击【确定】按钮，则以后所有的分析将仅仅基于符合选择条件的记录。

选择个案后的数据视图如图 3-45 所示, 不满足选择条件的个案其编号(如“1”、“2”、“3”, ...)被打上了反斜杠, 表示这些个案被过滤掉, 没被选中, 将不应用于分析; 而变量性别、种族和地区分别取值为“男性、黑种人、东北部地区”的个案将被选中, 其个案编号没有任何标记(如个案“6”、“7”)。同时 SPSS Statistics 根据选择条件生成了“filter_\$”筛选器变量, 如果其值为“Selected”表示该个案被选中, 如果其值为“Not Selected”表示该个案没有被选中。此外 SPSS Statistics 右下角的“筛选范围”表示该数据文件已被成功执行过“选择个案”的操作。

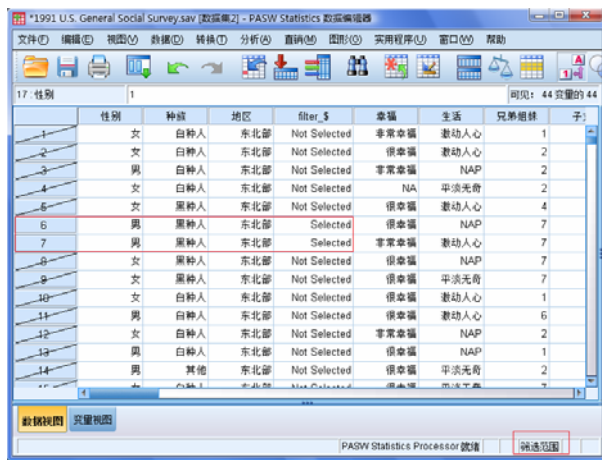


图 3-45 选择个案后数据视图

生成“filter_\$”变量选择个案的语法如下:

```
COMPUTE filter_$=(性别 = 1 & 种族 = 2 & 地区 = 1).
VARIABLE LABEL filter_$ '性别 = 1 & 种族 = 2 & 地区 = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

“选择个案”除了以上根据特定的条件进行选择外, 还有其他三种选择个案的方法:

1. 随机选择个案 (如图 3-46 所示)

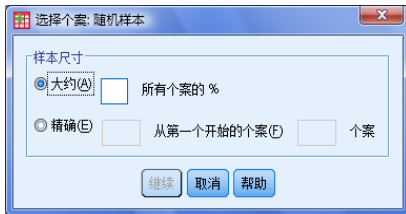


图 3-46 随机选择个案

在“大约 (A)”后的框中输入所选个案的百分比，将实现随机抽取该百分比的个案数。

选择“精确 (E)”，并在其后的文本框中输入具体的个案数 5，在“从第一个开始的个案 (F)”中输入 100，则表示在第 1 至第 100 个案中，随机选择 5 个个案。

2. 基于记录号选择个案（如图 3-47 所示）

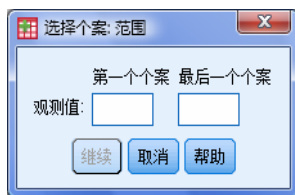


图 3-47 基于记录号选择个案

在“观测值”后的文本框中输入个案的范围，比如输入 5 和 100，则表示选择了第 5 至第 100 个个案，总共 96 个个案。

3. 使用筛选器变量选择个案（如图 3-48 所示）

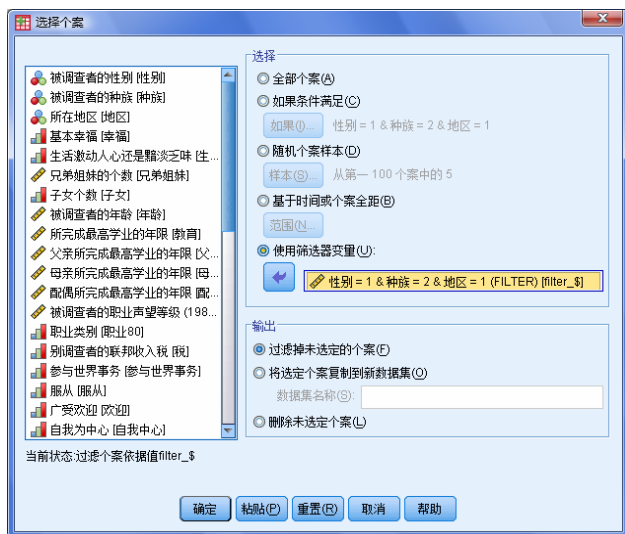


图 3-48 使用筛选器变量选择个案

将“筛选器变量 filter_\$”选入“使用筛选器变量 (U)”，即可实现之前定义的“男性、黑种人、东北部地区”个案选择。输出选择个案的方式，如图 3-49 所示。

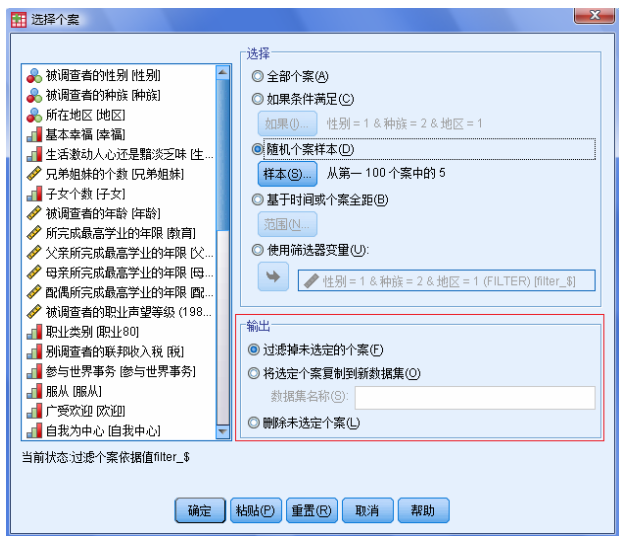


图 3-49 输出选择个案的方式

在图 3-49 “输出” 部分中各选项的作用解释如下。

- “过滤掉未选定的个案”：该选项为默认选项，不删除未选定的个案，只是过滤掉未选定的个案，未选定个案仍在数据集中，该效果如之前所述。
- “将选定个案复制到新数据集”：将对选定的个案生成一个新的数据集，数据集的名称由使用者自己命名。
- “删除未选定个案”：在当前数据集中删除未选定个案而只留下选定个案，该选项不推荐使用，除非使用者非常确信以后不再需要分析未选定个案。

综上所述，“选择个案”功能可实现对特定个案的选择，并基于所选定个案进行统计分析和建模。

注意：在应用选择个案完成所需要的分析后，立即关闭选择个案是一个好的习惯。否则，后续的分析将仅仅基于上一次所选择的个案。

3.6 小结

本章主要讲述了 SPSS 进行数据预处理的几种方法：可视化变量分段是对连续数据进行离散化，它把变量取值的细节进行简化、分类，它是对总体进行把握的一种十分有用的方法；缺失值填补是 SPSS 提供给数据分析工作者的一种简单易行的处理缺失数据的方法；而数据校验可以方便地让分析者找到现有数据中不一致或者录入数据中存在的错误。标识异常个案可以快速侦测数据中潜在的错误，它可以找出由于录入导致的数据错误，也可以找出不合常规的异常个案。最后介绍了选择符合指定条件的部分个案的方法。

思考与练习

1. 对 1991 U.S. General Social Survey.sav 进行个案选择, 选择条件为“女性、白种人、生活为平淡无奇”的个案, 并统计这些个案的年龄和教育的平均值和标准差。
2. 一家保险公司想找出那些可以的, 具有潜在骗保的客户的索赔个案。他们以前的索赔数据存储在数据文件“索赔数据.sav”中。但是由于他们的人力和财力有限, 不可能逐一对索赔客户进行一一的调查验证。因此, 在用当前数据建立模型以前, 他们计划用 SPSS 的自动数据准备功能, 在任何数据转换实际应用以前, 他们想先观测一下这种转换可能结果。因此, 请应用 SPSS 的自动数据准备工具, 通过交互式的方式演示他们所有可能采取的转换的潜在效果。另外, 找到重复进行索赔的客户的名单; 定义可疑客户规则, 并据此规则找到可疑的客户。
3. 在 SPSS 中有几种不同的方式进行个案选择, 其中正确的是:
 - A) 首先创建一个过滤器变量, 然后用 SELECT IF 语句根据过滤器变量进行选择
 - B) 【数据】→【选择个案】, 然后输入选择条件
 - C) 在 SPSS 数据编辑器中直接选择符合条件的个案
 - D) 在 SPSS 数据编辑器的变量视图中直接选择符合条件的个案
4. 应用 SPSS 的选择记录 (Select Cases) 菜单, 我们可以:
 - A) 选择符合指定逻辑条件的记录
 - B) 随机选择一定比例的记录
 - C) 从数据文件中删除某些记录
 - D) 添加某些符合条件的记录
5. 哪些方式是 SPSS 缺失值的替代方式:
 - A) 序列均值
 - B) 临近点的均值
 - C) 临近点的中位数
 - D) 线性插值法
 - E) 上一个记录的取值
6. 哪种方式可以关闭选择个案:
 - A) 运行【数据 (D)】→【选择个案】, 然后选择“全部个案 (A)”

- B) 在相应的数据视图中删除相应的筛选器变量
 - C) 直接删除个案编号前的反斜杠
 - D) 关闭数据文件
7. 有关 SPSS 选择某些特定条件的个案的说法, 不正确的是:
- A) 用菜单【数据(D)】→【选择个案】进行选择时, 没有被选中的个案可以仍然保留在数据集中
 - B) 用菜单【数据(D)】→【选择个案】进行选择时, 没有被选中的个案将不再保留在数据集中
 - C) 在应用 **SELECT IF** (逻辑表达式) 选择个案时, 逻辑表达式部分没有必要应用过滤器变量
 - D) 用菜单【数据(D)】→【选择个案】进行选择时, 没有被选中的个案将不再保留在数据集中
8. 在异常个案分析中, 对于 SPSS 报告的异常个案, 正确的处理方式:
- A) 在进行数据分析前, 应该首先删除 SPSS 探测出的异常个案
 - B) 异常个案探测是在数据分析完成之后, 进入结果验证阶段所做的分析
 - C) 即使合理地设置了异常探测的选项, 对于 SPSS 给出的异常个案也应该根据数据的具体含义和特点有区别的对待, 不能全部从数据集中删除
 - D) 在进行异常个案分析前, 应该先对缺失值进行填补
9. 有关 SPSS 的数据校验过程, 正确的是:
- A) 不用任何指导, SPSS 软件可以对数据进行校验, 找出出错的数据
 - B) 用户必须首先定义数据规则, 然后才能应用 SPSS 数据校验过程
 - C) 数据校验过程可以找出数据集中的所有错误
 - D) 数据校验过程必须对所有变量逐个进行

参考文献

1. 《SPSS 18 数据校验模块白皮书》。
2. 《SPSS 初中级培训讲义》。

描述性统计分析

本章学习目标：

- 掌握对数据进行描述的图形化方法和数值方法；
- 学习分析数据分布的方法；
- 掌握应用 SPSS 进行描述性数据分析的方法；
- 掌握常用统计图形的绘制方法和解释技巧。

统计分析的目的是研究观察对象总体的特点。在现实生活中，一般无法得到观察对象的总体，只能从总体中抽取一部分，我们称这部分为一个样本。统计学就是通过样本数据来研究总体数据的一门学科。它可以分为描述性统计分析和推断性统计方法。描述性统计方法是指应用分类、制表、图形以及概括性数据指标（例如均值、方差等）来概括数据分布特征的方法。而推断性统计方法则是通过随机抽样，应用统计方法把从样本数据得到的结论推广到总体的数据分析方法。例如，分析某个钢铁公司最近 5 年的经营状况，可以通过条形图、均值、方差等描述性统计方法；但是，不能够把从分析该公司得到的结论推广到所有的钢铁公司。而推断性统计，则是应用随机抽样的方法，抽取许多家钢铁公司，然后应用 T 检验，卡方检验等方法来分析得到的结果是由于抽样的偶然性还是普遍存在的。推断性统计分析得到的结论适用于总体。本章介绍在 SPSS 软件中进行描述性统计分析的方法。第 5 至第 9 章则介绍推断性统计分析的方法。

统计分析往往是从了解数据的基本特征开始的。统计上，需要把样本数据所含信息进行概括、融合和抽象，从而得到反映样本数据的综合指标，这些指标称为统计量。描述数据特征的统计量可分为两类：一类表示数据的中心位置，例如均值、中位数、众数等；另一类表示数据的离散程度，例如方差、标准差、极差等用来衡量个体偏离中心的程度。两类指标相互补充，共同反映数据的特征。

在进行数据分析时，第一步往往是先进行描述性统计分析，对数据做出大致的

判断，为以后对总体进行正确统计推断打好基础。

4.1 频率分析

在描述定性观测值时，有时候我们需要把这些值按照某种原则分成一些组或者类，并且使得每个观测值必须落入一个类并且只能够落入一个类中。对于给定的类，落入这个类的个案数称为频率，落入该类中的个案数和个案总数的比例称为相对频率。频率分析主要通过频率分布表、条形图和直方图，以及集中趋势和离散趋势的各种统计量来描述数据的分布特征。

打开本章示例数据 **Employ Data.sav**，该数据记录了某公司职工的基本信息，例如性别、民族、出生日期、教育水平、工资水平、工作年限等。教育水平为分类变量，它有 11 个类别。我们下面对“教育水平”变量进行频率分析以了解该公司员工的受教育水平。

选择【分析】→【描述统计】→【频率】，出现如图 4-1 所示的“频率分析”对话框，把“教育水平”变量选入右侧的“变量(V)”框中。



图 4-1 “频率分析”对话框

在图 4-1 下方，如果勾选“显示频率表格”复选框，可在输出中显示统计变量各具体值的频率、百分比、有效百分比、累计百分比，并且显示统计变量的有效和无效的个案数，输出结果如图 4-6 所示。如果不勾选“显示频率表格”复选框，则分析结果仅仅显示统计变量的有效和无效的记录数，如图 4-5 所示。我们采用默认值，要求输出频率表格，则得到输出结果如图 4-5 和图 4-6 所示。

在图 4-1 右方，单击【统计量(S)】按钮，出现图 4-2 所示对话框，从中可以选择需要的统计量。它们分别是描述性统计分析指标：百分位值、集中趋势、离散和

分布。4.2 节将介绍集中趋势；4.3 节介绍离散和百分位值（即分位数）；4.4 节介绍分布。

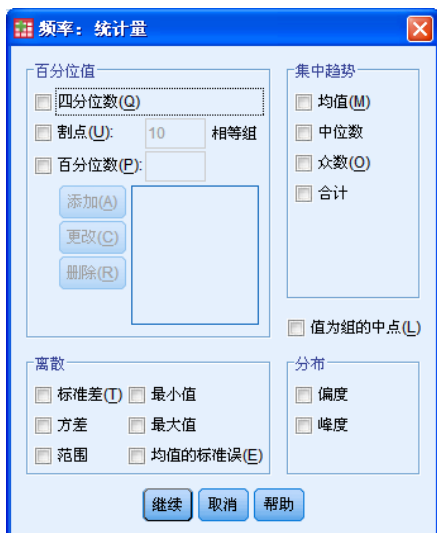


图 4-2 频率统计量对话框

在图 4-1 右方，单击【格式(F)】按钮，出现图 4-3 所示对话框，它将设置频率表输出的排序方式。如果选择“按值的升序排列”或者“按值的降序排列”，则频率表将按照个案值的升序或者降序排列；如果选择“按计数的升序排序”或者“按计数的降序排序”，则频率表将按照各个类别的频率值进行升序或者降序排列。

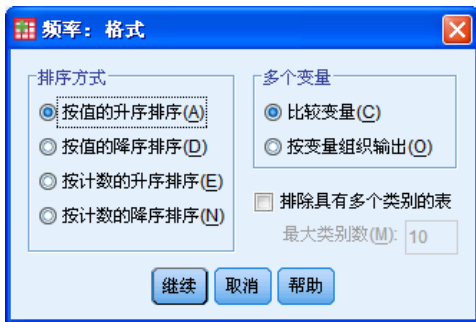


图 4-3 频率表格式对话框

在图 4-1 右方，单击【图表(C)】按钮，出现如图 4-4 所示的“频率：图表”对话框，供用户选择图形方式来描述数据，可供选择的统计图有条形图、直方图和饼图。输出结果除了图 4-5 和图 4-6 以外，还可以输出所选中的统计图。这里我们选择条形图，得到“教育水平”变量的频率分析结果如图 4-5 到图 4-7 所示。

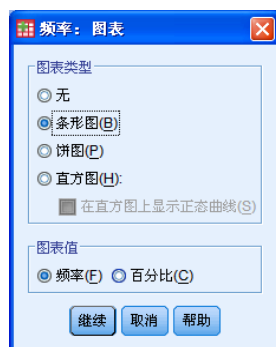


图 4-4 “频率：图表”对话框

统计量

教育水平（年）

N	有效	474
	缺失	0

图 4-5 频率分析结果 1

教育水平（年）

		频率	百分比	有效百分比	累积百分比
有效	8 年	53	11.2	11.2	11.2
	12 年	190	40.1	40.1	51.3
	14 年	6	1.3	1.3	52.5
	15 年	116	24.5	24.5	77.0
	16 年	59	12.4	12.4	89.5
	17 年	11	2.3	2.3	91.8
	18 年	9	1.9	1.9	93.7
	19 年	27	5.7	5.7	99.4
	20 年	2	.4	.4	99.8
	21 年	1	.2	.2	100.0
	合计	474	100.0	100.0	

图 4-6 频率分析结果 2

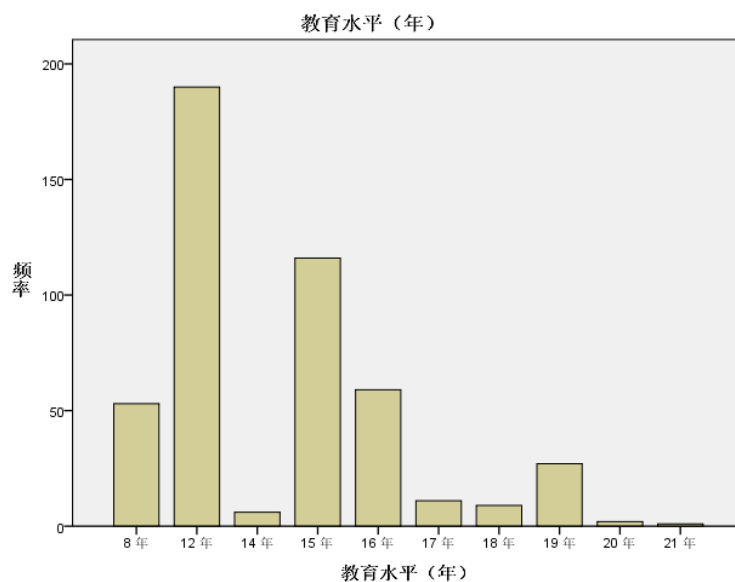


图 4-7 频率分析——条形图

注意：在频率分析中，饼图和条形图一般适用于分类变量的类别个数较少的情况，如果类别个数较多，例如多于 10 类，建议选择直方图。

4.2 中心趋势的描述：均值、中位数、众数、5%截尾均值

中心趋势是指一组数据向某个中心值靠拢的倾向。在统计学中，描述数据分布的中心位置的统计量称为位置统计量。对于连续变量（或称为尺度变量）和定序变量，描述数据中心趋势的指标有均值、中位数、众数、5%截尾均值；对于定性数据（即名义数据），描述数据中心趋势的指标只有众数。

注意：SPSS 中把变量分为三个水平，尺度变量、定序变量、名义变量。在 SPSS 变量编辑窗口要恰当的定义变量的水平，这是选择正确统计方法的基础。

有时候称尺度变量数据为连续数据，称名义变量数据为定性数据。统计学上把名义变量和定序变量统称为分类变量。

4.2.1 均值

均值即数据的算术平均数，是数据中心趋势的主要度量指标，也是实际问题中使用最多的指标。设我们考察的变量有 n 个测量值，它们分别记为 x_1, x_2, \dots, x_n ，则算术均值为：

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

数据的均值容易受极端值的影响，例如考察下列两组观测值：

组 1：1, 3, 5, 7, 9

组 2：1, 3, 5, 7, 14

两组观测值除了最后一个值以外全部相同，即其中心趋势应该大致相同。由于组 2 的最后一个值远远大于其他取值，导致组 2 的均值大于组 1 的均值。这两组的均值示例图如图 4-8 所示。

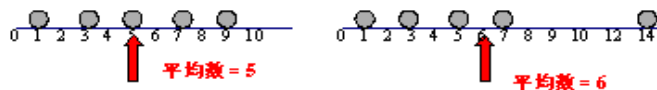


图 4-8 均值示例图

4.2.2 中位数

将观测值按照从小到大的顺序排列，位于中间位置的数值称为中位数。可以在中位数位置把数据分成两部分，一部分大于该数值，一部分小于该数值，这两部分各占观测值个数的百分之五十。在相对频率直方图中，一半的面积位于中位数位置的左边，一半的面积位于中位数位置的右边。

设我们考察的变量有 n 个测量值，它们分别记为 x_1, x_2, \dots, x_n ，把它们按照从小到大的顺序排列，排序后的数值记为 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 。

如果 n 为奇数，则中位数为 $M = x_{\frac{n+1}{2}}$

如果 n 为偶数，则中位数为 $M = \frac{x_{n/2} + x_{n/2+1}}{2}$

中位数受极端值的影响较小，在具有极大或极小值的数据中，中位数比均值往往更能代表数据的集中趋势。

中位数适合任意分布的数据，由于中位数只是考虑到居中位置，其他变量值相对于中位数的大小则无法反映。所以，用中位数来描述连续变量时会损失很多信息。当样本比较小时，中位数不太稳定，并不是一个很好的选择。但对于对称分布的数据，统计上往往会选择用中位数来描述数据的集中趋势。

在 4.2.1 中的例子，虽然两组数据不同，第二组观测值有极端值，但是它们的中位数相同，都是 5。

4.2.3 众数

众数是观测值中出现次数最多的数值，它反映了这组观测值的集中趋势。例如，调查十个学生的统计学成绩，它们的成绩分别为：

69, 72, 84, 75, 84, 75, 74, 89, 90, 75

这组数据中 75 分出现了 3 次，84 分出现了 2 次，其他成绩都只出现一次。因此，75 是众数。

又例如，在某个交通路口观测来往车辆 2 小时，共通过 1459 动车，其分布情况为：

小轿车 卡车 大客车 拖车

因此，通过车辆的众数为小轿车，即经常看到的是小轿车。

注意：1.众数是定性数据仅能使用的中心趋势指标，但众数可以用于尺度数据。
2.众数不一定唯一，甚至有时候众数不存在。

4.2.4 5%截尾均值

某些比赛是集体评分。先由每个裁判给出评分，然后去掉最高评分和最低评分，剩余得分的均值作为最终的得分。把观测值按照从小到大顺序排列，剔除掉排序后的数据序列两端的部分数值后计算得到的均值，称为截尾均值。SPSS 的描述性分析提供 5%截尾均值，它是把观测值升序排列后，剔除掉最小的 5%和最大的 5%后的数据的算术均值。这样计算出的均值就避免了极端值的影响。

假设我们考察的变量有 n 个测量值，它们分别记为 x_1, x_2, \dots, x_n ，把它们按照从小到大的顺序排列，排序后的数值记为 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 。

另外，记 $tc = 0.05n$ ，求得 k_1 和 k_2 使它们满足 $k_1 - 1 < tc < k_1$ ， $n - k_2 < tc < n - k_2 + 1$ ，

则 5%截尾均值为：

$$T = \frac{1}{0.9n} (k_1 - tc)y_{k_1} + (n - k_2 + 1 - tc)y_{k_2} + \sum_{i=k_1+1}^{k_2-1} y_i$$

4.3 离散趋势的描述：极差、方差、标准差、分位数和变异指标

仅仅根据数据的中心趋势指标进行决策是不够的。例如，如果一个国家的不同家庭收入差距很少；而另一个国家的家庭收入差距很大，既存在大量的贫困家庭，也存在许多十分富有的家庭，那么即使这两个国家的中等收入家庭的收入完全一样，其家庭收入情况仍然完全不同。对于一种药而言，如果一些批次的活性成分浓度太高，而其他批次活性成分浓度太低，那么即使这种药的活性成分平均浓度是正确的，该药物仍然十分危险。

假设我们有以下的三组观测值：

观测 A：11, 12, 13, 16, 16, 17, 18, 21

观测 B：14, 15, 15, 15, 16, 16, 16, 17

观测 C: 11, 11, 11, 12, 19, 20, 20, 20

这三组观测值的均值都是 15.5，亦即他们的中心趋势指标一样，那么这三组数据是否相似呢？从如图 4-9 所示的中心趋势与离散趋势看出，它们偏离中心的情况是完全不一样的。观测 B 的观测值最集中，观测 A 的观测值相对较分散，而观测 C 的观测值则偏离中心最大。

由此可见，仅仅了解数据的集中趋势是不够的，我们还需要了解数据波动范围的大小。描述数据波动大小的指标即为离散趋势指标。常用的离散趋势指标有：全距、方差、标准差、变异系数。

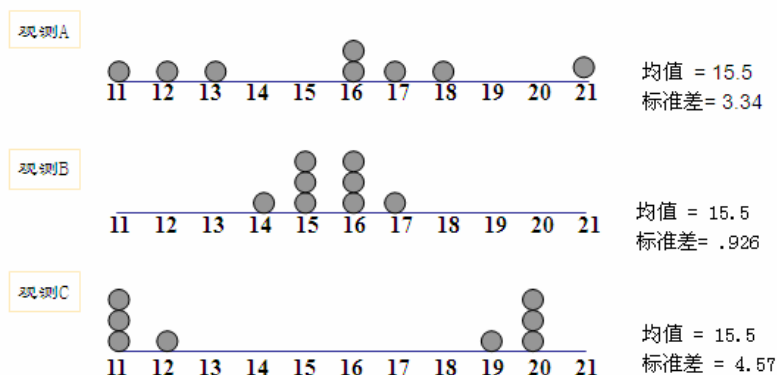


图 4-9 中心趋势与离散趋势

4.3.1 极差 (Range)

全距又称极差，是观测值中最大值与最小值之差。极差反映了变量的变异范围或离散幅度，任何两个观测值的差距都不会超出全距。全距仅仅由观测值的两个极端值确定，没有充分利用全部观测数据，它容易受极端值的影响。在 SPSS 中文版中，Range（极差）被译为“范围”。

4.3.2 方差和标准差

对每个观测值而言，其离散程度的大小就是其偏离均值的情况，即该观测值与均值的差值。这个差值可以用来描述个体的变异大小，但它不能表示整体的离散程度。因为所有数据与均值的差值之和正好是零，因此需要采用差值的绝对值之和来衡量整体偏离均值的大小。由于绝对值在数学上处理不是很方便，因而采用等价的差值平方和。即计算每个观测值与均值的差值平方，然后把所有平方值相加，这就

是方差。

假设我们考察的变量有 n 个测量值，它们分别记为 x_1, x_2, \dots, x_n ，其均值为 \bar{x} ，则方差为：

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

方差在使用上有一点不方便，就是量纲不合常理。例如身高变量的量纲为米，则其方差的量纲就是平方米了，这显然与其实际意义不符。因此将方差开平方，就得到了标准差。标准差的计算公式为：

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

标准差用来度量观测值偏离平均数的大小，相当于平均偏差，可以直接描述数据偏离均值的程度。对于同质的数据，一个较大的标准差，代表大部分的数值和其平均值之间差异较大；一个较小的标准差，代表这些数值较接近平均值。

4.3.3 变异系数

假设我们收集了 50 个职工的两个变量值，一个变量是平均收入（单位：美元），而另一个变量为员工的高度（单位：米）。假设二者方差相同，都是 28。那么，我们是否可以说明员工的高度变量和平均收入变量的波动程度相当呢？这要取决于比较的这两个变量的量纲，假设员工平均收入为 \$19700，平均高度为 1.71 米，那么我们可以知道对于员工平均收入，相对于其接近 2 万的均值，方差 28 实在算不得太大，而对于身高则相反。

可见在比较两组数据离散程度大小时，如果数据的测量尺度相差太大，或者是数据的量纲不一样，这时直接比较二者的标准差并不合适。需要首先消除测量尺度和量纲的影响。变异系数就可以剔除这些影响，其计算公式为：

$$V_\sigma = \frac{\bar{x}}{s}$$

4.3.4 分位数

分位数又称为百分位数，它是一种位置指标。 $p\%$ 分位数是指使得至少有 $p\%$ 的

数据小于或等于这个值,且使得至少有 $(100-p)\%$ 的数据大于或等于这个值。 $p\%$ 分位数位置的计算公式为 $i=(p/100)\times n$,即将数据按照从小到大进行排序,第 i 个位置的数即为 $p\%$ 分位数。前面所讲到的中位数,就是第50百分位数。

除了百分位数外,还有四分位数和十分位数。四分位数就是将观测数值按从小到大进行排序,然后分成四等份,处于三个分割点位置的观测值就是四分位数。最小的四分位数称为下四分位数,记为 Q_1 。所有观测值中,有四分之一的观测值小于下四分位数,四分之三大于下四分位数。中点位置的四分位数就是中位数。最大的四分位数称为上四分位数,记为 Q_3 。所有观测值中,有四分之三的观测值小于上四分位数,四分之一大于上四分位数。下四分位数有时也叫第25百分位数或第一个四分位数;而上四分位数也叫第75百分位数或第三个四分位数。

实际中,通常用 Q_3 和 Q_1 的差值来衡量观测值的离散程度,即四分位距:

$$IQR = Q_3 - Q_1$$

注意: 总结五数 (Five Number Summary) 和箱图 (Box Plot)

统计中常常把数据的最小值、下四分位数、中位数、上四分位数和最大值称为数据的总结五数。从这五个值可以大致看出数据分布的中心和离散程度。而箱图则是这五个数的图形表现,具体参见4.6.2节。

4.4 分布的形状——偏度和峰度

集中趋势和离散程度是数据分布的两个重要特征,但要全面了解数据分布的特点,还需要掌握数据分布的情况,例如其分布图形是否对称、偏斜程度以及扁平程度等。反映这些分布特征的统计指标是偏度和峰度。偏度 (Skewness) 用来描述变量取值分布的偏斜方向,它衡量分布对称与否、分布不对称的方向和程度。样本的偏度系数为:

$$\alpha = \frac{1}{ns^3} \sum_{i=1}^n (x_i - \bar{x})^3$$

α 的取值一般在-3和3之间。当 $\alpha > 0$ 时,分布为正偏或右偏,即分布图形在右边拖尾,如图4-10(A)所示,分布图有很长的右尾,尖峰偏左; $\alpha < 0$,分布为负偏或左偏,即分布图形在左边拖尾,如图4-10(B)所示,分布图有很长的左尾,峰尖偏右; $\alpha = 0$,分布对称。不论正、负哪种偏态,偏度的绝对值越大表示偏斜的程度越大;反之偏斜程度越小,分布形状越接近对称。

如图 4-10 (A) 和图 4-10 (B) 所示的两幅直方图可以看出, 图 4-10 (A) 是左偏, 图 4-10 (B) 是右偏。它们的偏度分别为-0.83 和 2.125。

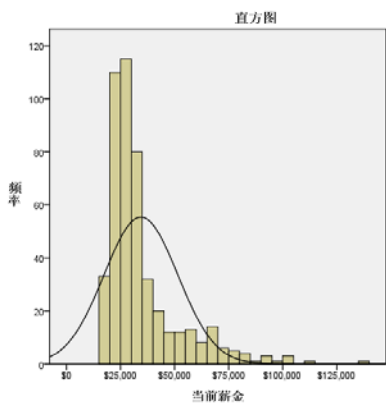


图 4-10 (A) 偏度-右偏

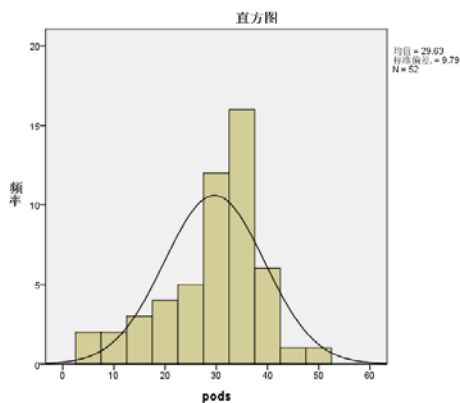


图 4-10 (B) 偏度-左偏

峰度是用来描述变量取值分布形态陡缓程度的统计量, 是指分布图形的尖峭程度或峰凸程度。样本的峰度系数为:

$$\beta = \frac{1}{ns^4} \sum_{i=1}^n (x_i - \bar{x})^4$$

$\beta > 3$, 分布为高峰度, 即比正态分布的峰要陡峭; $\beta < 3$, 分布为低峰度, 即比正态分布的峰要平坦些; $\beta = 0$, 分布为正态峰。

在 SPSS 中, 峰度计算公式是对公式 2 变化后的。如果 SPSS 给出的峰度值为 0, 分布为正态峰; 如果峰度值为负值, 为低峰度, 观测值在分布中心附近没有正态分布那样集中, 尾部更厚; 如果峰度值为正值, 则为尖峰, 即和正态分布相比, 有更多的观测值聚集在分布的中心位置, 尾部更薄。

4.5 SPSS 描述性统计分析

SPSS 的许多菜单均可进行描述性分析, 许多统计过程也都提供描述性统计指标的输出。例如在独立样本 T 检验、方差分析、因子分析等许多分析过程中, 都在结果中提供相应变量的均值、标准差等统计量。另外, SPSS 自定义表模块也可以产生大部分的描述性统计指标。

专门为描述性统计分析而设计的几个菜单集中在【分析】→【描述统计】菜单中, 如图 4-11 所示, 最常用的是列在最前面的四个过程:

- 频率 (F)：该过程将产生频数表；
- 描述 (D)：该过程则进行一般性的统计描述；
- 探索 (E)：该过程用于对数据概况不清时的探索性分析；
- 交叉表 (C)：该过程完成分类数据的统计描述和一般的统计检验，我们常用的 χ^2 检验也包含在如图 4-11 所示的描述性统计分析菜单中。



图 4-11 描述性统计分析菜单

我们以 Employ Data.sav 数据的“当前薪金”变量为例，讲解 SPSS 进行描述性统计分析的方法。

4.5.1 频率入口

在 SPSS 中选择【分析】→【描述】→【频率】，出现如图 4-12 所示频率对话框。前面 4.1 节已经介绍了频率表格，我们这里不再显示该表格。在图 4-12 中去掉对话框下部“显示频率表格”前面的钩。



图 4-12 频率对话框

然后单击【统计量】按钮。得到如图 4-13 所示的频率：统计量对话框。该对话框中的“百分位值”部分设置是否输出四分位数、是否输出十分位数和需要输出的

百分位数，默认输出四分位数和十分位数。“集中趋势”部分设置需要输出的描述集中趋势的统计量。“离散”部分设置需要输出的离散趋势统计量。“分布”部分设置是否输出偏度或者峰度。

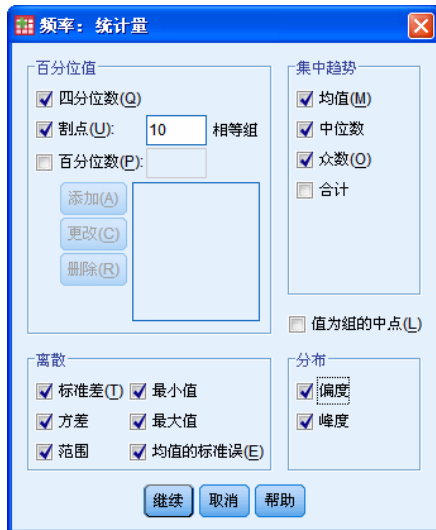


图 4-13 设置输出统计量

按照图 4-13 所示设置相应的选项，然后单击【继续】按钮。得到频率中的统计量输出结果（经过编辑处理），如图 4-14 所示。

统计量		
当前薪金		
N	有效	474
	缺失	0
均值		\$34,419.57
均值的标准误		\$784.311
中值		\$28,875.00
众数		\$30,750
标准差		\$17,075.661
方差		291578214.453
偏度		2.125
偏度的标准误		.112
峰度		5.378
峰度的标准误		.224
全距		\$119,250
极小值		\$15,750
极大值		\$135,000
百分位数	10	\$21,000.00
	20	\$22,950.00
	25	\$24,000.00
	30	\$24,825.00
	40	\$26,700.00
	50	\$28,875.00
	60	\$30,750.00
	70	\$34,500.00
	75	\$37,162.50
	80	\$41,100.00
	90	\$59,700.00

图 4-14 频率中的统计量输出

4.5.2 描述子菜单

在 SPSS 中选择【分析】→【描述统计】→【描述】，得到如图 4-15 所示的“描述性”对话框，把“当前薪金”选入变量框中。



图 4-15 “描述性”对话框

然后单击【选项】按钮。得到如图 4-16 所示的“描述选项”对话框，这里可以设置需要输出的描述性统计分析指标。

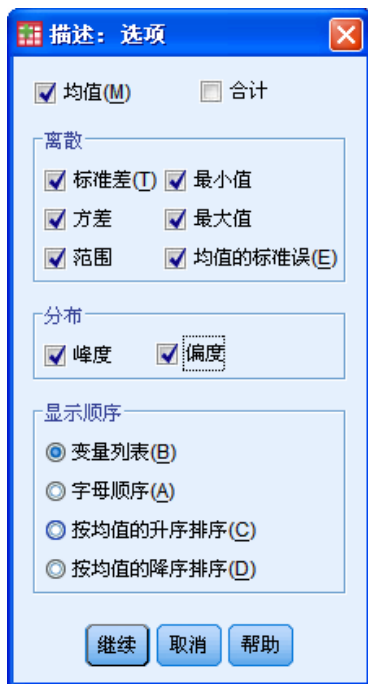


图 4-16 “描述选项”对话框

按照图 4-16 设置相应的选项，然后单击【继续】按钮，得到如图 4-17 所示的

描述统计量结果。

描述统计量		当前薪金	有效的 N (列表状态)
N	统计量	474	474
全距	统计量	\$119,250	
极小值	统计量	\$15,750	
极大值	统计量	\$135,000	
均值	统计量	\$34,419.57	
	标准误	\$784.311	
标准差	统计量	\$17,075.661	
方差	统计量	2.916E8	
偏度	统计量	2.125	
	标准误	.112	
峰度	统计量	5.378	
	标准误	.224	

图 4-17 描述统计量结果

4.5.3 探索子菜单

在 SPSS 中, 选择【分析】→【描述统计】→【探索】, 得到如图 4-18 所示的“探索”对话框。把“当前薪金”选入到因变量列表部分。然后在对话框下部的“输出”部分中, 我们仅选“统计量”。



图 4-18 “探索”对话框

设置如图 4-18 所示，然后单击“统计量(S)”按钮，得到如图 4-19 所示的统计量设置对话框。我们选择在描述性、界外值和百分位数这三个统计量前打钩。

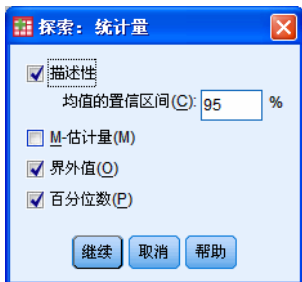


图 4-19 统计量设置

单击【继续】按钮，回到上级页面（探索对话框）后单击【继续】按钮，得到“当前薪金”变量的描述统计量、百分位数、最高和最低的 5 个极值等分析结果，如图 4-20 到图 4-22 所示。

描述			统计量	标准误
当前薪金	均值		\$34,419.57	\$784.311
	均值的 95% 置信区间	下限	\$32,878.40	
		上限	\$35,960.73	
	5% 修整均值		\$32,455.19	
	中值		\$28,875.00	
	方差		2.916E8	
	标准差		\$17,075.661	
	极小值		\$15,750	
	极大值		\$135,000	
	范围		\$119,250	
	四分位距		\$13,163	
	偏度		2.125	.112
	峰度		5.378	.224

图 4-20 描述统计量结果

百分位数		
百分位数	加权平均(定义 1)	Tukey 的枢纽
	当前薪金	当前薪金
5	\$19,200.00	
10	\$21,000.00	
25	\$24,000.00	\$24,000.00
50	\$28,875.00	\$28,875.00
75	\$37,162.50	\$37,050.00
90	\$59,700.00	
95	\$70,218.75	

图 4-21 百分位数

极值			
		案例号	值
当前薪金	最高	1	29
			\$135,000
		2	32
			\$110,625
		3	18
	最低		\$103,750
		4	343
			\$103,500
		5	446
			\$100,000
	最低	1	378
			\$15,750
		2	338
			\$15,900
		3	411
			\$16,200
		4	224
			\$16,200
		5	90
			\$16,200

图 4-22 极值

4.5.4 表格

在 SPSS “设定表” 菜单中也可以输出大部分的描述性统计分析指标。选择【分析】→【表】→【设定表】，得到如图 4-23 所示的“设定表格”对话框。用鼠标把需要分析的变量拖放到中间框中的行（W）框上。

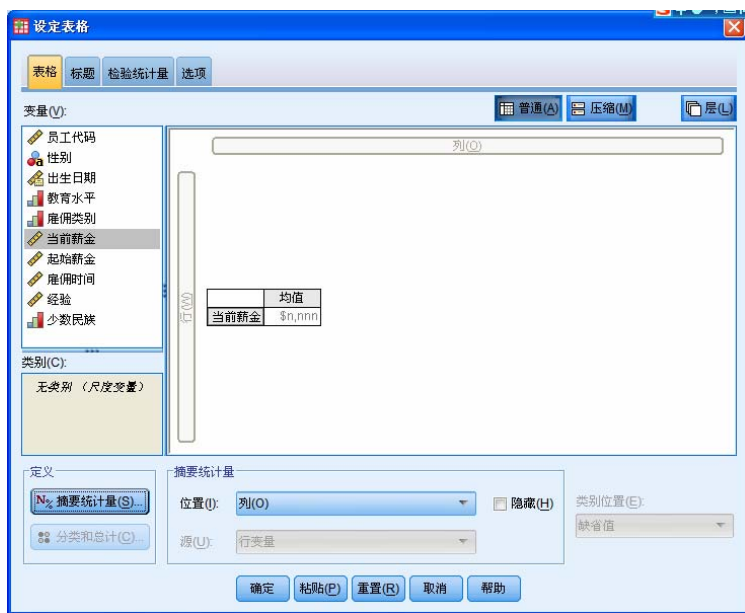


图 4-23 “设定表格”对话框

然后单击左下角的“摘要统计量（S）”按钮，弹出如图 4-24 所示的“摘要统计量”对话框，供用户选择需要输出的统计指标，例如均值、中位数、众数、方差和标准差等。

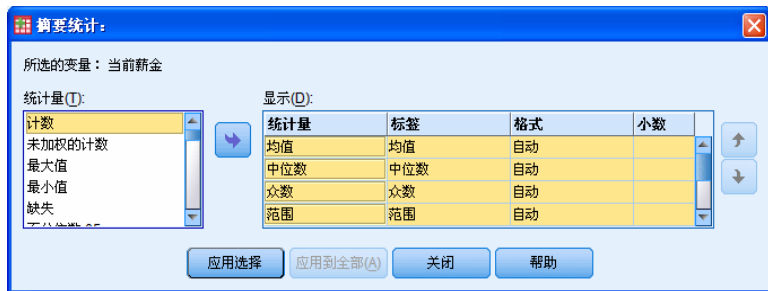


图 4-24 选择统计量

这里我们选择“均值”、“中位数”、“众数”、“范围”、“方差”、“标准差”、“百分位数”等，然后单击“应用选择”按钮，得到如图 4-25 所示的“设定表格”对话框，以设置预览。



图 4-25 设置预览

单击【确定】按钮，得到客户表输出结果，如图 4-26 所示。

	均值	极大值	中值	极小值	众数	范围	均值的标准误	标准差
当前薪金	\$34,420	\$135,000	\$28,875	\$15,750	\$30,750	\$119,250	\$784	\$17,076

图 4-26 客户表输出结果

4.6 应用统计图进行描述性统计分析

描述性统计分析除了应用数量指标以外，还可以应用条形图、饼图、帕累托图、直方图、箱图、茎叶图等统计图形，相应的统计图选项分布在【图形】菜单或者某些分析过程的相应选项下。本节主要介绍在输出描述性统计量的同时，可以选择的统计图形。在【分析】→【描述统计】→【频率】子菜单下的“图表”选项，可以选择绘制条形图、饼图和直方图。

- 在【分析】→【描述统计】→【探索】子菜单下的“绘制”选项，可以绘制箱图、茎叶图、直方图、和检验数据正态性的 Q-Q 图，并且可以选择是否按照分组来绘制箱图。

注意：除帕累托图位于【分析】菜单的【质量控制】子菜单以外，所有的统计图都可以在 SPSS 的【图形】菜单下得到。

一个好的习惯是，在进行统计分析前，总是把数据“画出来”，即做出数据的相关的统计图

4.6.1 定性数据的图形描述

定性数据的图形描述常用条形图、帕累托图或饼图表示。

- 条形图给出相应每一类的频率（或者相对频率），长方形的高度（注：水平方向条形图为长方形的长度）与类的频率或者相对频率成比例。
- 帕累托图是按照从高到低顺序排列条形图的长方形条后形成的一种特殊条形图，最高的长方形在左边。它是质量控制中常用的一种图形工具，其中长方形的高度通常表示生产过程中产生问题（如缺陷、事故、故障和失效）的频数，而最左边的长方形对应于最严重的问题区域。帕累托图形就是在【分析】菜单的【质量控制】子菜单下“排列图”。
- 饼图把一个整圆（饼）分成几份，每一份代表一个类，每份中心角与类相对频率成比例。

表 4-1 汇总了自 1977 年以来全世界 45 起与能源有关导致多人死亡的事故的原因。该数据显然是定性数据，它保存在本章的数据文件 DisasterReason.sav 中。

表 4-1 与能源有关的导致多人死亡事故的原因的相对频率汇总表

类（原因）	频率（事故数）	相对频率（比例）
煤矿坍塌	7	0.156
溃坝	4	0.089
煤气爆炸	28	0.622
闪电	1	0.022
核反应堆	1	0.022
燃气火灾	4	0.089
总计	45	1.000

由于数据提供的是汇总后的频数数据，在应用 SPSS 绘制统计图之前，先用频数来加权个案。选择【数据】→【加权个案】，得到如图 4-27 所示的“加权个案”对话框。

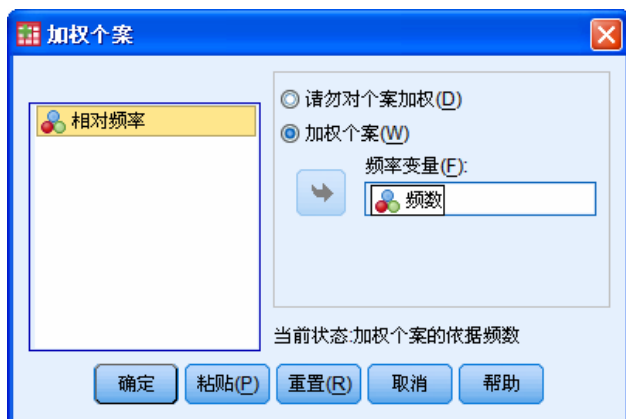


图 4-27 “加权个案”对话框

把“频数”作为频率变量，单击“确定”按钮，返回到 DisasterReason.sav 数据集，现在数据是已经加权后的数据。

单击【分析】→【描述】→【频率】得到如图 4-28 所示的“频率”对话框，去掉底部“显示频率表格”前的勾，然后单击【图表】按钮，得到如图 4-29 所示的“频率：图表”对话框，以选择图表类型。



图 4-28 在“频率”对话框中选择输出统计图

在图 4-29 中，我们选择条形图，图表值是设置长方形高度所代表的指标，这里默认为频率。请读者自己练习选择饼图。

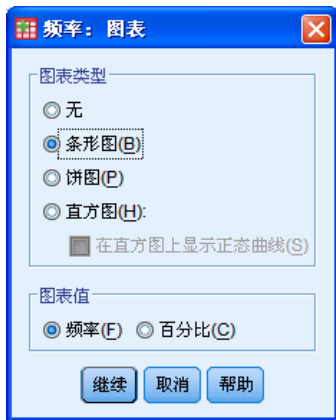


图 4-29 选择图表类型

单击【继续】按钮，返回到如图 4-28 所示的频率对话框，在其中选择输出统计图的页面，然后单击【确定】按钮。得到如图 4-30 所示的事故原因条形图和如图 4-31 所示的饼图。

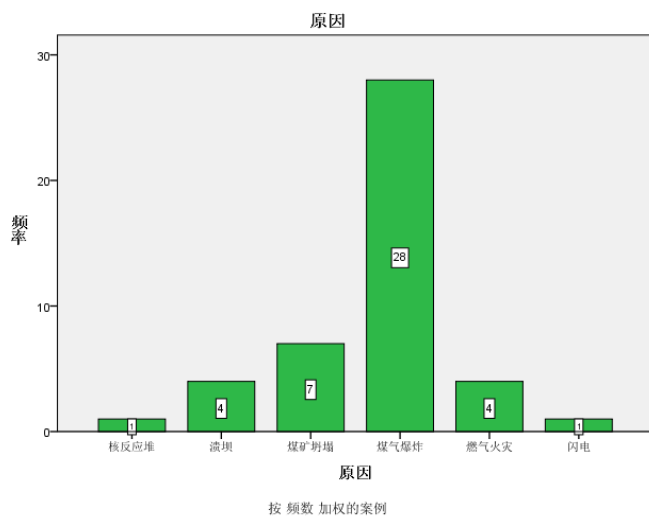


图 4-30 事故原因条形图

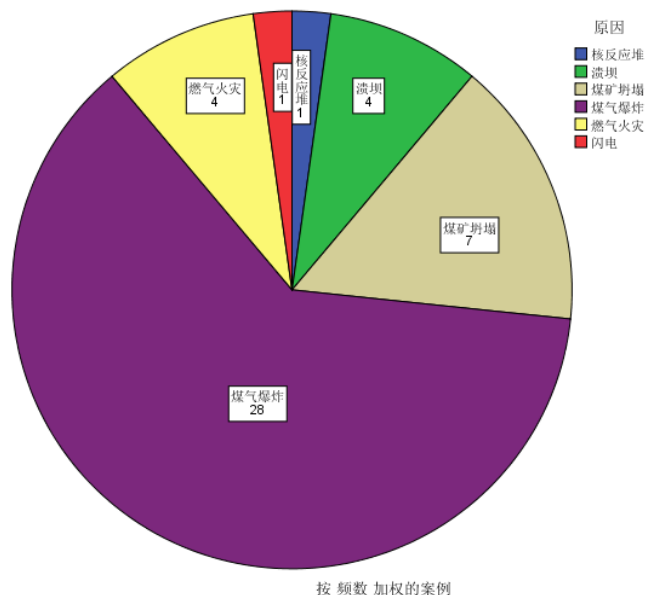


图 4-31 事故原因饼图

帕累托图实质上是按照长方形条从高到低排序的条形图，从它可以一眼看出煤气爆炸是最可能引起事故的原因，它对应排在最左边的最高的长方形条。下面示例做出表 4-1 的帕累托图，选用上面加权过的数据文件，选择【分析】→【质量控制】→【排列图】，得到如图 4-32 所示的“帕累托图”数据选择对话框。

注意：如果直接应用未加权的数据文件 DisasterReason.sav，图表中的数据部分可以选择“个案值”选项。

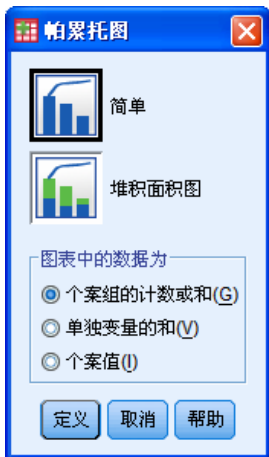


图 4-32 “帕累托图”数据选择对话框

在图 4-32 中选择第一个图标—简单，图表中的数据部分选择“个案组的计数或和”，单击“定义”按钮。得到如图 4-33 所示的定义帕累托图对话框。



图 4-33 定义帕累托图对话框

把“原因”变量选入“类别轴（X）”框中，单击“确定”按钮。得到如图 4-34 所示的帕累托图，图形上面的折线为各种原因的累积百分比，从上面可以看出各种原因在总的事故中所占的百分比。

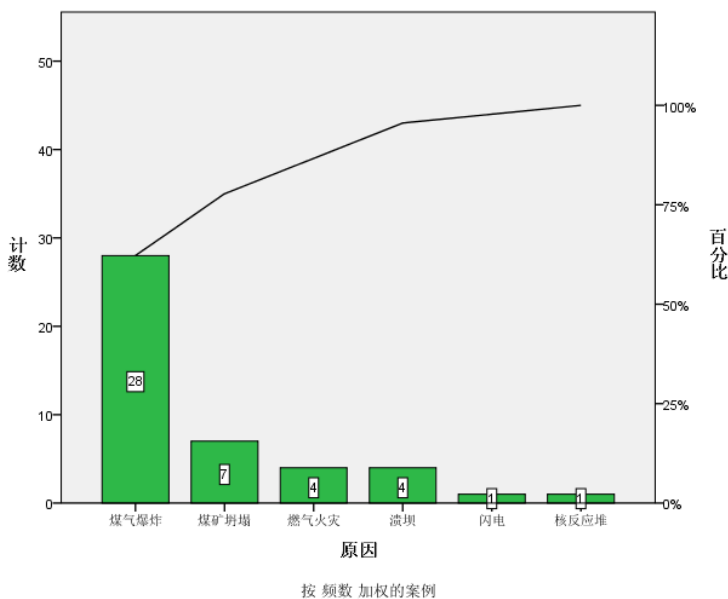


图 4-34 带累积线的帕累托图

4.6.2 定量数据的图形描述

定量数据可以采用三种统计图形来描述：直方图、茎叶图和箱图。

1. 直方图

直方图和条形图十分类似，它应用于连续型数据，表现在图形上直方图的各个正方形条之间没有任何间隔。它先把连续型数据划分成若干个连续的区间，然后计算观测值落入各个区间的频率或者相对频率。和条形图类似，以区间类作为水平轴，以各个区间的频率作为相应长方形的高度绘制出的统计图。

从直方图可以直观的观测数据的分布情况，例如分布是否对称、是左偏还是右偏、众数是什么。另外，还可以大致判断数据是否服从正态分布。

打开本章的示例文件 **Employ Data.sav**，我们来绘制“起始薪金”变量的直方图。与 4.6.1 节绘制条形图类似，在图 4-28 频率对话框中，变量部分选择“起始薪金”，单击“图表”按钮，得到如图 4-29 所示的选择图表类型对话框，这里选择“直方图”。

得到的直方图如图 4-35 所示。

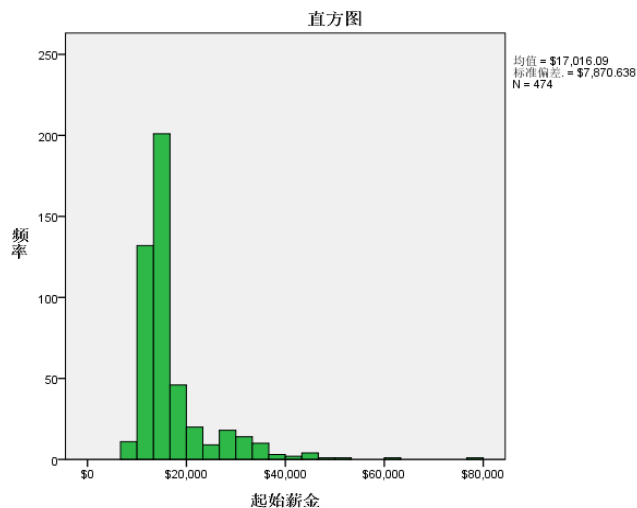


图 4-35 直方图

2. 茎叶图 (Stem-and-Leaf Plot)

茎叶图是描述定量变量的一种图形方式，它除了能够给出直方图所给出的分布的信息以外，还能够还原大部分原始数据的信息。

打开本章的示例文件 `PovertyByState.sav`，它记录了 1997 年统计的美国 51 个州处于贫困线以下的人口占各州人口的比例。在数据视图中的数据如图 4-36 所示。

	州	百分比
1	Alabama	14.80
2	Alaska	8.50
3	Arizona	18.80
4	Arkansas	18.40
5	California	16.80
6	Colorado	9.40
7	Connecticut	10.10
8	Delaware	9.10
9	D.C.	23.00
10	Florida	14.30
11	Georgia	14.70
12	Hawaii	13.00
13	Idaho	13.30
14	Illinois	11.60
15	Indiana	8.20
16	Iowa	9.60
17	Kansas	10.40
18	Kentucky	16.40
19	Louisiana	18.40
20	Maine	10.70
21	Maryland	9.30
22	Massachusetts	11.20
23	Michigan	10.70

图 4-36 数据视图

我们用 SPSS 绘制变量“百分比”的茎叶图。选择【分析】→【描述统计】→【探索】，在如图 4-37 所示的“探索”对话框下部的“输出”部分，选择“图”。



图 4-37 “探索”对话框——茎叶图

按照图 4-37 设置后，单击【绘制】按钮，得到如图 4-38 所示的选择统计图形对话框，选择统计图形的类别。箱图部分可选择箱图的类型。描述性设定输出茎叶图或者直方图。如果勾选“带检验的直方图”，可以输出选定变量的 Q-Q 图、变量正态性的 K-S 检验和 S-W 检验，据此来判断该变量是否服从正态分布。

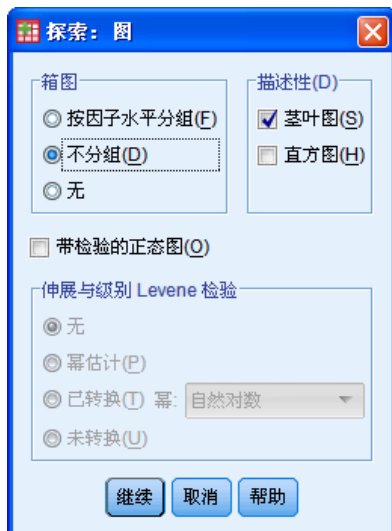


图 4-38 选择统计图形

单击【继续】按钮，返回上级，探索对话框——茎叶图对话框，然后单击【确定】按钮。得到茎叶图如图 4-39 所示。

百分比 Stem-and-Leaf Plot

Frequency	Stem & Leaf
1.00	0 . 7
11.00	0 . 88889999999
15.00	1 . 000000001111111
6.00	1 . 222333
6.00	1 . 444455
6.00	1 . 666667
4.00	1 . 8888
2.00	Extremes (>=23)
Stem width:	10.00
Each leaf:	1 case(s)

图 4-39 茎叶图

SPSS 输出的茎叶图由三部分构成：频率、茎和叶。茎对应观测值的最左边一位的取值，而叶对应最左边第二位的取值，在 Leaf 部分每一个数字代表一个个案。相应行左边的 Frequency 是该行对应的个案个数，即该分支中的个案的个数。

在图 4-39 中，最后一行的 Each Leaf: 1 case(s)意味着每一个个案对应一个叶节点，Stem Width:10 意味着茎是取观测值十位数上的值，如果观测值小于 10，则相应的茎为 0。第一行告诉我们有一个个案其百分比在 7%和 8%之间（这里对应 New hampshire 州）。第二行意味着有 11 个个案的百分比在 8%和 10%之间，其中有 4 个州的百分比在 8%和 9%之间，有 7 个州的百分比在 9%到 10%之间。

3. 箱图 (Box Plot or Box-and-Whisker Plot)

箱图是总结五数（最小值、第一个四分位数、中位数、第三个四分位数、最大值）的图形表现。箱图在比较两组或者两组以上的观测值时尤其有用，另外它也可以用于判断离群值（或者极端值）。

在图 4-38 选择统计图形对话框中，在箱图部分，选择相应的箱图类型。当具有一个或多个因变量时，这些选项控制箱图的显示。

- 按因子水平分组：如果在图 4-37 的“因子列表 (F)”部分设定了因子变量，则为每个因变量生成单独的显示。在一个显示中，将为因子变量定义的每个组显示箱图。
- 不分组：如果在图 4-37 的“因子列表 (F)”部分设定了因子变量，将按照因子变量定义的每个组生成单独的显示。在一个显示中，为每个因变量

并排显示箱图。当不同的变量代表在不同的时间度量的同一个特征时，此显示尤其有用。

- 无：不输出箱图。

在图 4-38 选择统计图形中，我们设置输出箱图，结果浏览器中得到如图 4-40 所示的箱图。

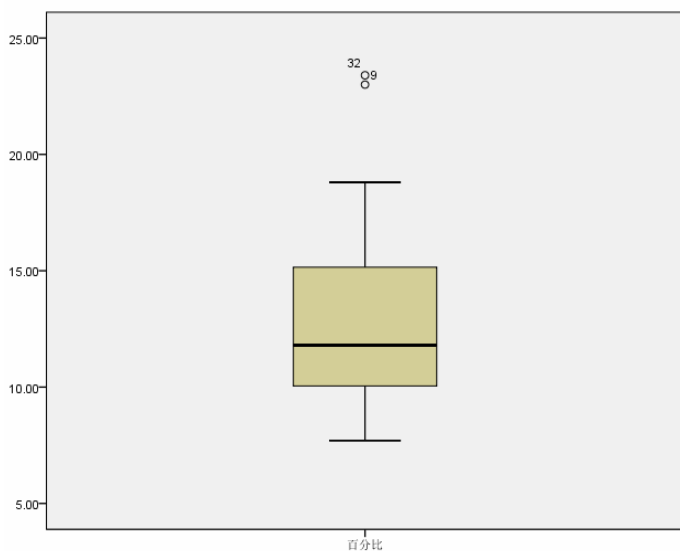


图 4-40 箱图

设四分位距为 $IQR = Q_3 - Q_1$ 。箱图的箱体部分的下边界代表第一个四分位数的位置，上边界代表第三个四分位数的位置，中间的粗体线段代表中位数的位置，箱体的高度即为四分位距 IQR 。最下面的短线代表 $Q_1 - 1.5IQR$ 的位置，最上面的短线代表 $Q_3 + 1.5IQR$ 的位置。

如果观测值落入 $[Q_3 + 1.5IQR, Q_3 + 3IQR)$ 或者 $(Q_1 - 3IQR, Q_1 - 1.5IQR]$ ，则该观测值为离群值，在箱图上用小圆圈标识，并在它的旁边给出该个案的记录号；如果观测值大于等于 $Q_3 + 3IQR$ 或者小于等于 $Q_1 - 3IQR$ ，则该观测值被判为极端值，在箱图上用星号标识，并在其旁边给出该个案的记录号。

在贫困百分比变量的箱图中，有两个离群值，他们的记录号为 9 和 32，分别对应华盛顿州 D.C 和新墨西哥州 New Mexico。

如果只有一个因变量，茎叶图或者箱图将按照因子的各个水平输出，选择“不分组”的输出结果和选择“按因子水平分组”的输出结果只是在标题的组织方式上

略有不同，如果有两个因变量，则两种选项的结果差异较大。如果描述性分析同时有两个因变量，因子有两个水平，在箱图选项中选择“按照因子水平分组”，则输出如图 4-11 所示。如果选择“无分组”，则结果如图 4-42 所示。

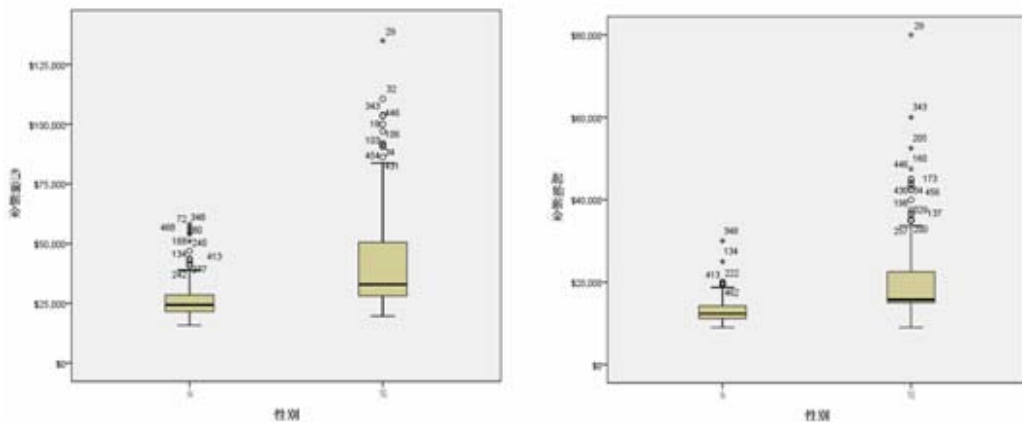


图 4-41 按因子水平分组

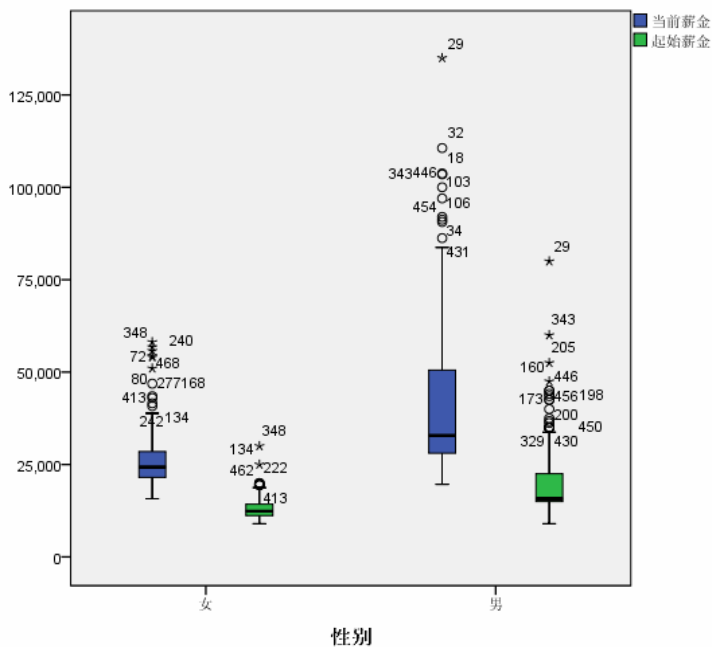


图 4-42 不分组

4.7 数据标准化

数据标准化处理主要包括数据同趋化处理和无量纲化处理两个方面。数据同趋

化处理主要解决不同性质数据问题，对不同性质指标直接加总不能正确反映不同作用力的综合结果，必须先考虑改变指标数据性质，使所有指标对测评方案的作用力同趋化，再加总才能得出正确结果。数据的标准化处理有很多种方法，比如 Z 标准化。标准化处理后，可以保证数据服从标准正态。具体计算公式为：

$$Z = \frac{x - \bar{x}}{s}$$

另外针对其他标准化方法，可以根据设定的公式，新建一个衍生变量。

在 SPSS 中，在【分析】→【描述统计】→【描述】中，可以选择进行数据标准化，结果会在原始的数据文件中，新生成一个变量，该标准化变量就是应用上面公式的算法，“数据描述性”操作如图 4-43 所示，数据标准化的结果如图 4-44 所示。

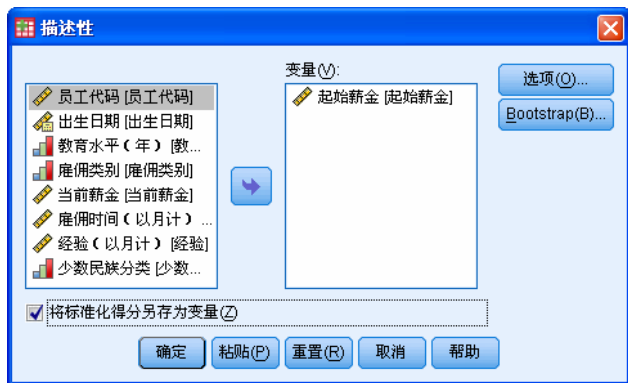


图 4-43 “数据描述性”对话框

	教育水平	雇佣类别	当前薪金	起始薪金	雇佣时间	经验	少数民族	Z起始薪金	变量	变量	变量	变量	变量	变量
1	15	3	\$57,000	\$27,000	98	144	0	1.36993						
2	16	1	\$40,200	\$19,750	98	36	0	.22030						
3	12	1	\$21,450	\$12,000	98	301	0	-.63732						
4	8	1	\$21,900	\$13,200	98	190	0	-.40495						
5	16	1	\$45,000	\$21,000	98	138	0	.80617						
6	16	1	\$32,100	\$13,500	98	67	0	-.44673						
7	15	1	\$36,000	\$19,750	98	114	0	.22030						
8	12	1	\$21,900	\$9,750	98	0	0	-.92319						
9	15	1	\$27,900	\$12,750	98	115	0	-.54203						
10	12	1	\$24,000	\$13,500	98	244	0	-.44673						
11	16	1	\$30,300	\$16,500	98	143	0	-.06557						
12	8	1	\$28,350	\$12,000	98	26	1	-.63732						
13	15	1	\$27,750	\$14,250	98	34	1	-.35144						
14	15	1	\$36,100	\$16,800	98	137	1	-.02745						
15	12	1	\$27,300	\$13,500	97	66	0	-.44673						
16	12	1	\$40,800	\$15,000	97	24	0	-.25615						
17	15	1	\$46,000	\$14,250	97	48	0	-.35144						
18	16	3	\$103,750	\$27,510	97	70	0	1.33330						
19	12	1	\$42,300	\$14,250	97	103	0	-.35144						
20	12	1	\$26,250	\$11,550	97	48	0	-.69449						
21	16	1	\$38,650	\$15,000	97	17	0	-.25615						
22	12	1	\$21,750	\$12,750	97	315	1	-.54203						
23	15	1	\$24,000	\$11,100	97	75	1	-.75167						
24	12	1	\$16,950	\$9,000	97	124	1	-1.01848						
25	15	1	\$21,150	\$9,000	97	171	1	-1.01848						

图 4-44 数据标准化结果

4.8 小结

本章主要介绍了描述性统计分析的方法和技巧。4.1 节介绍了分类变量的频率分析，4.2 节到 4.4 节分别介绍了常见的描述性统计分析指标。其中，描述中心趋势的指标有：均值、中位数、众数、5%截尾均值；描述离散趋势的指标有：极差、方差、标准差、变异系数；另外，还有描述数据的分布形态的指标，例如分位数、偏度和峰度。4.5 节介绍了如何在 SPSS 中得到描述性统计分析指标。4.6 节介绍了如何应用统计图对数据进行描述性统计分析。最后，介绍了在 SPSS 中对数据进行标准化的方法。

思考与练习

1. 在【图形】菜单中，重新做出 4.6 节的统计图形，比较这两种绘制统计图形的方法的异同点。
2. 指出均值、众数、中位数这三个描述数据中心趋势的指标有何区别，各有什么优缺点。
3. 说明茎叶图和直方图区别。如果想尽可能展现原始数据的信息，应该采用那一种图形？
4. 说明帕累托图和直方图的区别。
5. 请指出哪种衡量中心趋势的指标适宜用来描述下列属性，如果有两个以上的指标都可以，请指出哪个指标可以反映最多的信息量。
 - A) 姊妹和兄度的个数
 - B) 驾驶的汽车类型
 - C) 父亲的体重
 - D) 每年休假的天数
6. 对于上题中的四个变量，它们分别可以用哪种统计图形来描述？
 - A) 直方图
 - B) 条形图
 - C) 帕累托图
7. 一个数据文件包含下列数据：5 个家庭没有汽车（编码为 0）；20 个家庭拥有一辆汽车（编码为 1）；10 个家庭拥有两辆车（编码为 2）。指出下列哪

种统计量适用于描述该数据并计算出该统计量的值。

- A) 拥有汽车数的众数
- B) 拥有汽车数的中位数
- C) 拥有汽车数的方差
- D) 变异系数

8. 为了生成某个给定变量的总和（即“total”），应该选用哪一个汇总统计量：

- A) mean
- B) sum
- C) median
- D) mode

9. 假设有数据如图 4-45 所示，如果要求出 a, b, c 三个变量的均值，并且希望在有缺失值的情况下尽可能地利用已有数据的信息求出均值。在 SPSS 18 中，选择哪个函数可以达到要求：

CaseNo	a	b	c
1	0.5	0.6	0.7
2	0.3	0.2	
3		0.6	

图 4-45 数据集合图示

- A) mean(a,b,c)
- B) mean.2(a,b,c)
- C) mean2(a,b,c)
- D) (a+b+c)/3

10. 某公司的少数管理层员工有特别高的工资，大部分员工拿的工资很低。如果你代表员工去和公司老板谈判涨工资，那么你倾向于采用哪一个统计指标来说明员工的工资底；而如果你是老板，你倾向于采用哪一个统计指标来说明工人的工资已经很高了。

11. 箱图可以探测出数据中的异常值。请对数据 DisasterReason.sav 进行描述性统计分析，通过箱图分社数据中是否存在异常值。

参考文献

1. 梁冯珍 关静等译，《统计学（第 5 版）》，北京：机械工业出版社，2009。
2. David S. Moore, George P. McCabe, Introductory to the Practice of Statistics.

3. Michael Sullivan, III, Statistics, informed decisions Using Data, Prentice Hall.

均值的比较

本章学习目标：

- 掌握假设检验的基本思想；
- 掌握均值过程及其应用并能正确解释输出结果；
- 掌握单样本 T 检验的方法和应用条件，正确解释输出结果；
- 掌握独立样本 T 检验的方法和应用条件，正确解释输出结果；
- 掌握配对样本 T 检验的方法和应用条件，正确解释输出结果。

5.1 假设检验的思想及原理

假如你是某外贸公司的验货员，公司派你去供应商的工厂验货。考虑下列情形：

1) 该工厂的质量管理员跟你说：“根据我们的检验，我们的产品缺陷率只有千分之一，请检验吧！”。你接下来着手验货，从 1000 件产品中随机抽查了 5 件货品，发现其中 2 件货品都有质量缺陷。你的结论是什么呢？

如果工厂的质量管理员所说属实，则 1000 件产品中应该只有 1 件有缺陷产品，这称为原假设。随机抽查中不可能出现抽查到两件或者两件以上有缺陷的产品这一事件，即这一事件出现的概率为 0，在统计学中，概率极小的事件称为小概率事件。而现实是，随机抽查了 5 件产品，其中 2 件是有缺陷的。即小概率事件发生了。那么，我们只能说质量管理员的论断是错误的，抽查的结果不支持他的论断。

2) 如果该工厂的质量管理员跟你说：“根据我们的检验，我们的产品缺陷率只有百分之一，请检验吧！”。你接下来着手验货，从 1000 件产品中随机抽查了 5 件货品，发现其中 2 件货品都有质量缺陷。这时你的想法可能会有两种：

- 该工厂的货品缺陷率肯定远远高于 1%，达不到质量要求；
- 该工厂的货品缺陷率确实只有 1%，只是恰巧抽到有缺陷的产品。

那么哪种想法正确呢？这要比情形 1) 中难以判断。需要一些概率论的知识。

如果工厂的质量管理员所说属实,即该工厂产品缺陷率只有1%,也就是说假定原假设属实。则100件产品中应该只有1件有缺陷产品。那么1000件产品中有缺陷产品应该为10件。那么抽查5件产品,其中2件有缺陷这一事件的概率大约为

$$\frac{\binom{10}{2}\binom{990}{3}}{\binom{1000}{5}}=0.088\%$$

即该事件发生的概率小于万分之九,这可能性实在太小了,是小概率事件。

在原假设成立的条件下,如果计算出来样本所对应的事件发生概率比较大,那么没有理由拒绝原假设。反之,如果计算出来样本所对应的事件发生概率比较小,即小概率事件发生了,依据小概率事件在一次试验中是几乎不会发生的原理,它在一次实验中是不应该发生的。可事实是,本来不该发生的事件却在我们的试验中发生了。那么,我们只能说抽查结果不支持原假设中的论断,或者说质量管理员所声明的产品缺陷率只有1%的论断是错误的。

以上过程即整个假设检验的思想:反证法及小概率原理。所谓反证法及小概率原理即首先在原假设正确的条件下计算出出现该样本或者样本统计量的概率,如果这种事件发生的概率很小,譬如小于5%,那么就拒绝原来的假设,而接受备择假设(即工厂产品的缺陷率大于1%);如果这种事件发生的概率较大,譬如大于5%,则不推翻原假设。

尽管假设检验的依据是“小概率事件在一次试验中几乎不会发生”的原理,但是小概率事件并非是不可能发生,只是其发生的概率很小,我们并不能完全排斥其发生的可能性。因而假设检验有可能犯两类错误:

(1) 第一类错误:原假设正确,而错误地拒绝了它,即“拒真”的错误,其发生的概率为第一类错误的概率。在上面情形2)中,如果产品的真实合格率的确为1%,而我们认为产品合格率大于1%,则犯了该类错误,犯该类错误的概率为0.088%;

(2) 第二类错误:原假设不正确,而错误地没有拒绝它,即“受伪”错误,其发生的概率为第二类错误的概率。

在假设检验中,不可能同时降低犯第一类错误和犯第二类错误的概率。如果降低犯第一类错误的概率,则发生第二类错误的概率将提高,反之亦然。在上例中,工厂希望不要把合格的产品误判为不合格,即降低生产者的风险。实际中,犯第一

类错误受到更多的重视,希望把它控制在一定的水平,该水平称为显著性水平,用 α 表示。实际中经常取 0.05, 或者 0.01, 0.1 等。

假设检验一般先对总体的比例、均值或分布做出某种假设,称为原假设;然后计算在该假设成立条件下出现该事件的概率(或可能性),称为 p 值。如果小概率事件发生了,即 $p < \alpha$, 则表明样本不支持原来的假设,应拒绝原假设而接受备择假设;如果该事件发生的概率(或可能性)较大,即 $p > \alpha$, 则不拒绝原假设。我们用 α 来控制犯第一类错误的概率,即犯该类错误的概率最大为 α 。

假设检验的步骤为:

- 第一步: 确定恰当的原假设和备择假设;
- 第二步: 选择检验统计量;
- 第三步: 计算检验统计量观测值发生的概率,即 p 值;
- 第四步: 给定显著性水平 α , 并作出决策。如果 $p < \alpha$, 则拒绝原假设,反之,没有理由拒绝原假设。

实际中的许多假设检验问题都是比较两个总体的均值。均值的比较分析在实务中以实验研究最为常见,应用也最为广泛,如某些制药公司、食品公司以及许多公司的研发部门。

SPSS 比较两总体均值的方法全部都在【分析】菜单的【比较均值(M)】中,“比较均值”对话框如图 5-1 所示。

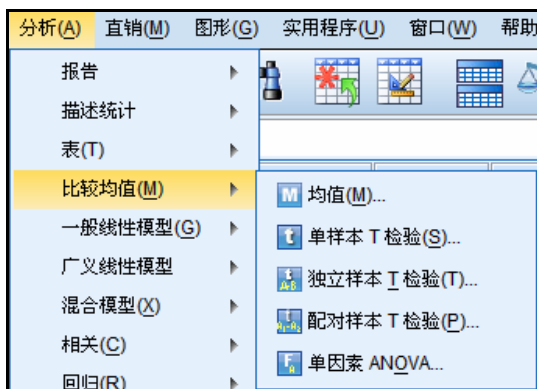


图 5-1 “比较均值”对话框

5.2 均值

SPSS 的均值过程是描述和分析尺度变量(Scale)的一种有用的方法,可以获

得需要分析的变量的许多中心趋势和离散趋势的统计指标，同时它可以对不同的组别或者交叉组别进行比较。该过程可以计算一个或多个自变量类别中因变量的子组均值和相关的单变量统计。也可以从该过程获得单因素方差分析、eta 和线性相关检验。例如，利用均值过程可以分析三类不同的烹调油所吸收的平均脂肪量，并执行单因素方差分析，查看均值是否相同。

从均值过程可以为每个分组变量的每个类别选择众多的子组统计量，它们包括：合计、个案数、均值、中位数、组内中位数、均值的标准误、最小值、最大值、范围、分组变量的第一个类别的变量值、分组变量的最后一个类别的变量值、标准差、方差、峰度、峰度标准误、偏度、偏度标准误、总和的百分比、总数的百分比、和的百分比、数量的百分比、几何均值以及调和均值。

本章的数据文件 HourlyWage.sav 是对护士工资的调查，它调查了不同位置的护士，记录了他们的小时工资、工作经验、年龄等指标。以下，应用 SPSS 的均值过程分析护士的小时工资、工作经验和工作位置之间的关系。

5.1.1 均值过程分析

打开数据文件 HourlyWage.sav，单击【分析】→【比较均值】→【均值】，出现如图 5-2 所示“均值”对话框。把 hourwage(小时工资)选入因变量列表中，把 yrsscale(工作经验) 选入自变量列表中。



图 5-2 “均值”对话框

单击【选项】按钮，选择统计量如图 5-3 所示。

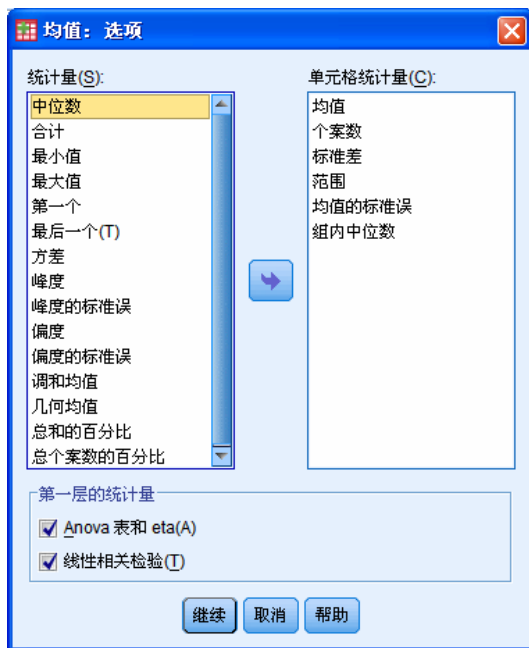


图 5-3 选择统计量

在图 5-3 中，可以选择需要的统计量。它们有中心趋势、离散程度和数据分布状况的各种描述性指标，可更改统计量出现的顺序。统计量在“单元格统计量 (C)”列表中出现的顺序是它们在输出中显示的顺序。还将显示跨所有类别的每个变量的摘要统计。这里，除了默认的均值、个案数和标准差以外，我们选择范围、均值的标准误和组内中位数。另外，在该对话框下面的“第一层的统计量”中，勾选“Anova 表和 eta (A)”和“线性相关检验 (T)”。单击【继续】按钮，返回如图 5-2 所示的对话框，然后单击【确定】按钮。

以上过程也可以通过以下语法命令完成：

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\ HourlyWage.sav'
DATASET NAME myData WINDOW=FRONT.
MEANS TABLES=hourwage BY yrsscale
    /CELLS MEAN COUNT STDDEV RANGE SEMEAN GMEDIAN
    /STATISTICS ANOVA LINEARITY.
```

输出中主要包括四部分：

1. 表 5-1 的“案例处理摘要”表给出分析中用到的有效个案数和排除的个案数（即含有缺失值的个案）。
2. 表 5-2 的“均值报告表”给出了因变量在各个子组的描述性统计量取值。表

5-3 的“方差分析表”给出了以工作经验为控制因素的方差分析表。

表 5-1 案例处理摘要

	案例					
	已包含		已排除		总计	
	N	百分比	N	百分比	N	百分比
小时工资 * 工作经验	2911	97.0%	89	3.0%	3000	100.0%

从表 5-1 知数据集中共有 3000 个案，其中 89 个案含有缺失值，均值过程是基于排除了含有缺失值个案后的 2911 个案。

表 5-2 均值报告表

小时工资

工作经验	均值	N	标准差	全距	均值的标准误	分组中值
少于 5 年	18.0416	221	3.86667	23.30	.26010	17.9342
6~10 年	18.9169	460	3.77816	24.72	.17616	19.0616
11~15 年	19.6616	752	3.90528	23.34	.14241	19.8514
16~20 年	20.2876	729	3.82786	23.57	.14177	20.6165
21~35 年	21.2594	539	4.08669	24.02	.17603	21.4741
36 年以上	21.6342	210	3.61826	21.26	.24968	21.5357
总计	20.0159	2911	4.00309	28.59	.07419	20.1800

均值报告表列出了各种工作经验的护士组别的小时工资均值、标准差、全矩、均值标准误，同时给出了不同工作经验的护士组别的个案数(N)和各组别小时工资的中位数（分组中值）。从均值列看出，随着工作年限的增加，小时工资也随之增加，但是增加的幅度不是均匀的。

“6~10 年”小时工资比“5 年以下”增加 0.88，随后，“11~15 年”和“16~20 年”比其前一级别增加的幅度变小，分别为 0.76 和 0.63，而“21~35 年”比“16~20 年”增加的幅度最大，为 0.97，之后“从 21~35 年”到“36 年以上”增加 0.34，其增加的幅度为最小。

表 5-3 方差分析表

			平方和	df	均方	F	显著性
小时工资 * 工作经验	组间	(组合)	2948.660	5	589.732	39.218	.000
		线性	2918.278	1	2918.278	194.070	.000
		线性偏差	30.381	4	7.595	.505	.732
	组内		43683.288	2905	15.037		
	总计		46631.948	2910			

3. 在表 5-3 的方差分析表中, 第二行“线性”的显著性值为“.000”(即 p 值小于 0.0005), 小于显著性水平 0.05, 可以判断小时工资和工作经验之间有线性关系; 第三行的“线性偏差”的显著性值为 0.732, 大于 0.05, 因此小时工资和工作经验之间的非线性关系的成分不显著。

表 5-4 相关性度量表

	R	R 方	Eta	Eta 方
小时工资 * 工作经验	.250	.063	.251	.063

4. 在如表 5-4 所示的相关性度量表中, R 方为 0.063, 该值不是太大。可以这样做出结论, 工作经验可以解释不同护士小时工资之间的差异, 但是工作经验和小时工资之间的线性关系不是十分强。

5.1.2 双因素的均值过程分析

在 5.1.1 中, 仅仅分析了影响小时工资的一个因素。考虑到除了工作经验不同之外, 护士的位置也可能会影响到小时工资。我们把护士的位置也作为均值过程的一个因素, 重新分析小时工资和护士的位置、工作经验之间的关系。

在图 5-2 中, 在“层 1 的 1”框中, 单击“下一张(N)”按钮, 出现如图 5-4 所示双因素的均值分析对话框。把 position 变量选入“层 2 的 2”框中的“自变量列表(I)”中, 其他保留默认值, 如图 5-4 所示。单击【确定】按钮。

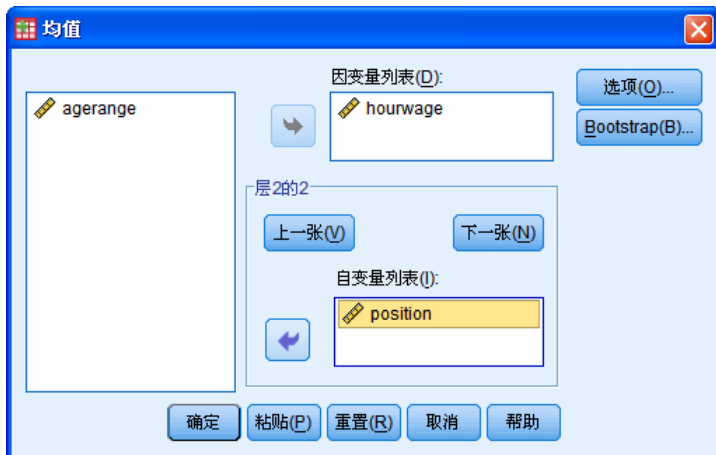


图 5-4 双因素的均值分析

以上操作过程可以用下列语法命令实现:

```

DATASET ACTIVATE myData.
MEANS TABLES=hourwage BY yrsscale

```


/CELLS MEAN COUNT STDDEV RANGE SEMEAN GMEDIAN
/STATISTICS ANOVA LINEARITY.

由于均值过程只对第一层的自变量进行方差分析和线性相关检验，因此两个因素或者两个以上因素的均值分析过程的方差分析结果和单因素一样。不同的是描述性统计量，多因素的描述性统计量是对于各个交叉组别进行统计，如表 5-5 所示。

表 5-5 两因素的均值分析报告

小时工资							
工作经验	护士类型	均值	N	标准差	全距	均值的标准误	分组中值
少于5年	病房	19.0753	147	3.37129	16.15	.27806	19.1847
	办公室	15.9882	74	3.98762	23.30	.46355	16.0356
	总计	18.0416	221	3.86667	23.30	.26010	17.9342
6-10年	病房	19.4846	313	3.35218	22.06	.18948	19.4737
	办公室	17.7082	147	4.32447	23.01	.35668	17.6434
	总计	18.9169	460	3.77816	24.72	.17616	19.0616
11-15年	病房	20.2412	518	3.41065	18.43	.14986	20.5259
	办公室	18.3784	234	4.57662	23.24	.29918	18.3900
	总计	19.6616	752	3.90528	23.34	.14241	19.8514
16-20年	病房	21.1369	471	3.29487	18.85	.15182	21.1777
	办公室	18.7373	258	4.23293	23.54	.26353	18.6611
	总计	20.2876	729	3.82786	23.57	.14177	20.6165
21-35年	病房	21.8601	350	3.48989	19.77	.18654	21.9363
	办公室	20.1471	189	4.82372	23.12	.35087	20.2619
	总计	21.2594	539	4.08669	24.02	.17603	21.4741
36年以上	病房	22.0641	146	3.14466	16.39	.26025	21.6995
	办公室	20.6534	64	4.38931	19.75	.54866	20.7361
	总计	21.6342	210	3.61826	21.26	.24968	21.5357
总计	病房	20.6764	1945	3.49582	24.16	.07927	20.7468
	办公室	18.6859	966	4.58852	27.69	.14763	18.6020
	总计	20.0159	2911	4.00309	28.59	.07419	20.1800

报告表的第一行是组别为“工作五年或者以下的病房护士”的描述性统计指标。第二行是组别为“工作五年或者以下的办公室护士”的描述性统计指标。第三行是合并以上两个组别，即组别为“工作五年或者以下护士”的描述性统计指标，它和表 5-2 的第一行一致。“均值”列显示各个交叉组别的小时工资的均值。从表 5-5 可看出，同等工作经验下，病房护士的小时工资比办公室护士工资高，但随着工作经验的增加，二者之间的差距变小。“标准差”列给出了各个组别小时工资的标准差，办公室护士的方差都要大于同等经验的医院护士，即同等经验的办公室护士，他们的小时工资差距要大于同等经验的医院护士。

5.3 单样本 T 检验

统计学的大部分理论是基于大样本的，即抽样的个体数比较大，个案数较多。

但是，样本量多大可以称为大样本？统计学上没有一个统一的规定，一些教材认为样本量 25 或者以上即为大样本，有些教材认为样本量 30 或者以上为大样本。在实际应用中，取决于分析数据的具体分布状况和模型的要求。

单样本 T 检验即检验某个变量的总体均值和某指定值之间是否存在着显著性差异。如果是大样本的单样本检验，统计教科书上称为 U 检验，它采用服从正态分布的 U 统计量作为检验统计量；如果是小样本并且样本服从正态分布，则采用服从 t 分布的 t 统计量进行单样本 T 检验；否则，采取非参数检验。T 检验稳健性 (Robust) 较好，如果样本分布偏离正态分布不太严重，也可采用 T 检验。

根据概率论中的中心极限定理，在大样本情况下，T 分布和正态分布密度函数十分接近，如图 5-5 所示。实际应用中，不论是大样本还是小样本，都可以用 T 检验来进行单样本均值检验。在大样本情况下，即使样本分布偏离正态的情况下，仍然可以应用 T 检验，这称为 T 检验的稳健性。因此，在大样本情况下，T 检验和 U 检验是等价的。

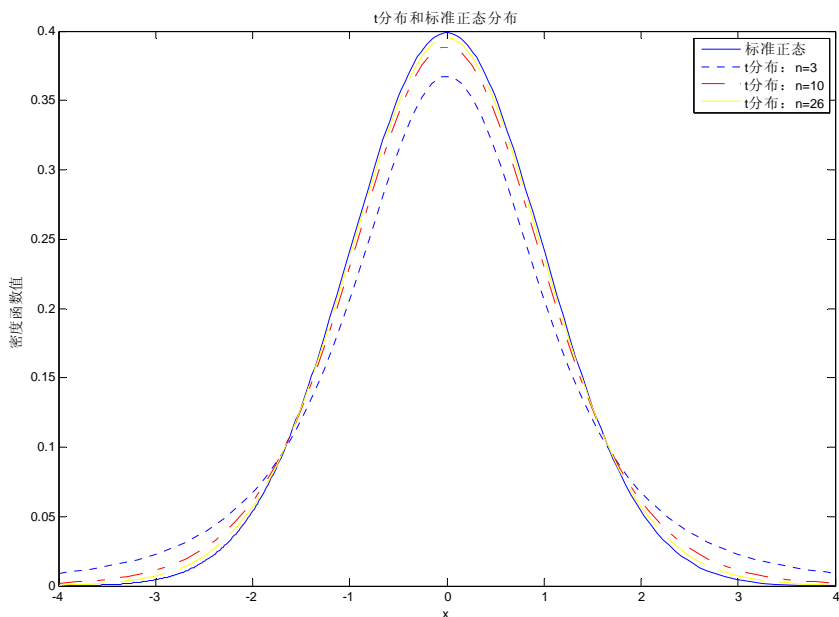


图 5-5 T 分布和正态分布的密度函数比较

打开本章数据文件 `brakes.sav`，该数据为某工厂不同机器生产的刹车片直径，已知符合质量标准的刹车片直径应为 322 mm，现在需要知道哪些机器生产的刹车片直径不符合质量标准（即对各个机器生产的刹车片直径进行总体均值为 322 的单样本 T 检验）。

5.3.1 数据准备

我们需要对各个机器分别进行检验，因此需要根据机器拆分该数据文件。打开数据文件 brakes.sav，选择【数据】→【拆分文件】，得到拆分文件菜单如图 5-6 所示。



图 5-6 拆分文件

在如图 5-7 所示的“分割文件”方式对话框的下面“当前状态：按组合分析关闭”意味着，当前的数据文件没有按组分割。选择“比较组”，把变量“机器”选入“分组方式 (G)”框中，单击【确定】按钮。这时，再重新打开分割文件对话框，下面将显示“当前状态：比较：机器”。即当前数据文件已经按照“机器”进行组织，基于该文件的分析将基于各个分组，直到关闭分割文件开关。

以上分割文件过程可以通过下列语法命令实现：

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\brakes.sav' .
DATASET NAME myData WINDOW=FRONT.
SORT CASES BY 机器.
SPLIT FILE LAYERED BY 机器.
EXECUTE.
```

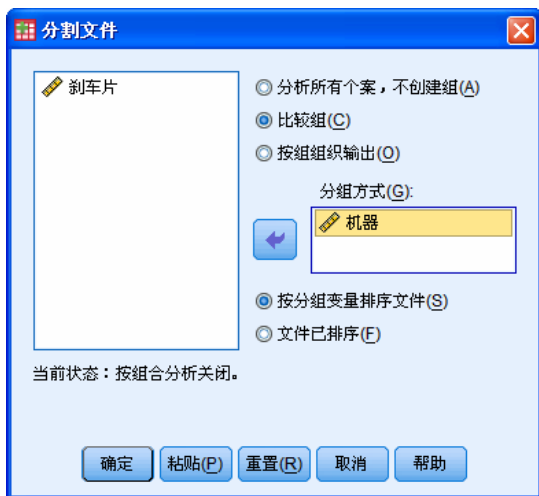


图 5-7 “分割文件”方式对话框

完成分析之后，可以运行下列语法命令关闭文件分割。

```

DATASET ACTIVATE myData.
SPLIT FILE OFF.

```

也可以通过如图 5-7 所示的对话框，选择第一个选项“分析所有个案，不创建组(A)”来关闭文件分割状态。

注意：当对文件进行分割，完成需要的分析之后，养成立即关闭文件分割的习惯。否则，下次打开文件会忘记当前文件的状态，后续的分析仍然基于分割后的数据。

5.3.2 单样本 T 检验

基于 5.3.1 分割后的数据文件，我们选择【分析】→【比较均值】→【单样本 T 检验】，如图 5-8 和图 5-9 所示，进入“单样本 T 检验”对话框。

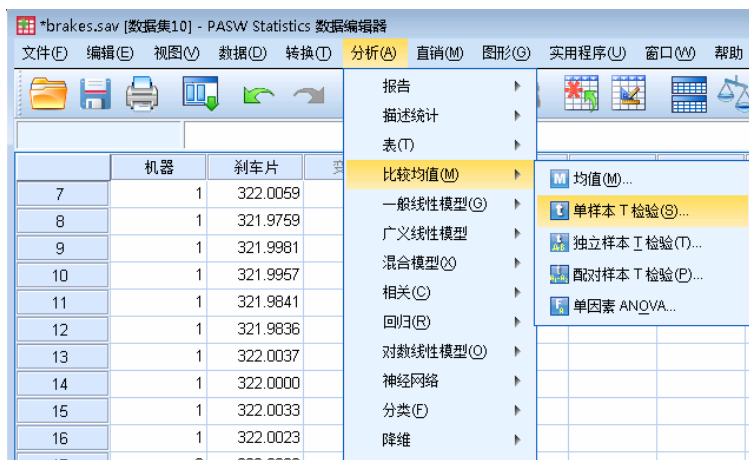


图 5-8 比较均值

在图 5-9 中，把“刹车片”选入“检验变量(T)”对话框中，并在“检验值(V)”框中输入待检验的总体均值“322”，其他保持不变，单击【确定】按钮。



图 5-9 “单样本 T 检验”对话框

以上操作过程可以通过下列语法命令实现：

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\brakes.sav' .
DATASET NAME myData WINDOW=FRONT.
DATASET ACTIVATE myData.
T-TEST
  /TESTVAL=322
  /MISSING=ANALYSIS
  /VARIABLES=刹车片
  /CRITERIA=CI(.95).
```

在输出窗口中得到分析结果如表 5-6 和表 5-7 所示。

表 5-6 单样本 T 检验描述性统计量

机器号	N	均值	标准差	均值的标准误
1 刹车片直径 (mm)	16	321.998514	.0111568	.0027892
2 刹车片直径 (mm)	16	322.014263	.0106913	.0026728
3 刹车片直径 (mm)	16	321.998283	.0104812	.0026203
4 刹车片直径 (mm)	16	321.995435	.0069883	.0017471
5 刹车片直径 (mm)	16	322.004249	.0092022	.0023005
6 刹车片直径 (mm)	16	322.002452	.0086440	.0021610
7 刹车片直径 (mm)	16	322.006181	.0093303	.0023326
8 刹车片直径 (mm)	16	321.996699	.0077085	.0019271

表 5-6 为各个机器号刹车片直径的均值、标准差及标准误统计量。从各个机器刹车片的均值来看，它们都或多或少偏离了 322 毫米。没有统计检验，很难判断哪个机器生产的刹车片直径不合格。

表 5-7 为单样本 T 检验的检验结果， t 列为 T 统计量的值， df 为自由度，Sig（双侧）为 p 值，“均值差值”为各个机器号的均值减去 322 的差，“下限”和“上限”列分别为该均值差的 95% 的置信区间的下限和上限。如果取 0.05 为显著性水平，如果 p 值小于 0.05 则通过显著性统计检验。从表 5-7 可知，机器 2、7 的 p 值（或称显著性值）都小于 0.05，且 t 值为正，因此它们的刹车片直径显著高于 322mm；机器 4 的刹车片直径显著低于 322mm；而机器 1、3、5、6、8 的 p 值都大于 0.05，因此不能说它们的刹车片直径不等于 322mm，也就是说这些机器号生产的刹车片直径满足生产质量要求。

表 5-7 单样本 T 检验结果

机器号	t	df	Sig.(双侧)	均值差值	差分的 95% 置信区间	
					下限	上限
1 刹车片直径 (mm)	-.533	15	.602	-.0014858	-.007413	.004459
2 刹车片直径 (mm)	5.336	15	.000	.0142629	.008566	.019960
3 刹车片直径 (mm)	-.655	15	.522	-.0017174	-.007302	.003868
4 刹车片直径 (mm)	-2.613	15	.020	-.0045649	-.008289	-.000841
5 刹车片直径 (mm)	1.847	15	.085	.0042486	-.000655	.009152
6 刹车片直径 (mm)	1.134	15	.274	.0024516	-.002154	.007058
7 刹车片直径 (mm)	2.650	15	.018	.0061813	.001210	.011153
8 刹车片直径 (mm)	-1.713	15	.107	-.0033014	-.007409	.000806

因此，得到的结论为，机器 2 和 7 生产的刹车片直径偏大，机器 4 生产的刹车片直径偏小。

5.3.3 置信区间和自抽样选项

在如图 5-10 所示的单样本 T 检验对话框中，可以设置在【选项】中定义均值的置信区间和设置自助法。

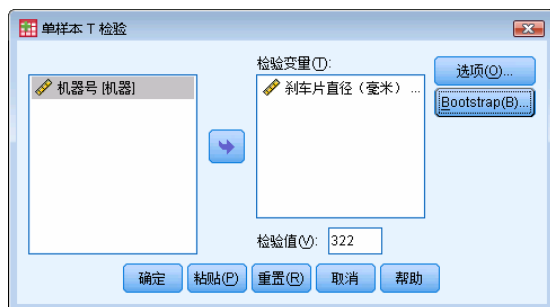


图 5-10 选项设置

单击【选项】按钮，可以更改置信区间的百分比，如图 5-11 所示。默认置信区间百分比为 95%，可以更改为其他需要的值，例如定义 90% 的置信区间。

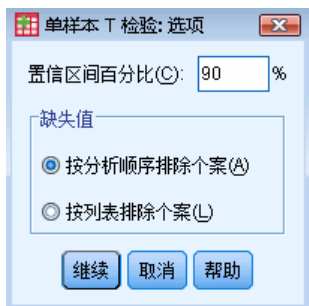


图 5-11 设置置信区间百分比

单击【Bootstrap】按钮可进行 Bootstrap 自抽样计算，用自助法进行置信区间的估计，设置自助的方法如图 5-12 所示，使统计量及假设检验更具稳健性和概化推论能力。



图 5-12 设置自助法

5.4 独立样本 T 检验

如果有两个或者以上的总体，需要考察这些总体间在统计学上是否有显著的区别。考察任何两个总体均值之间的区别，就是两样本的检验。如果考察三个或者以上总体均值之间的区别，则需要应用第 9 章中方差分析的技巧。从某种意义上而言，方差分析是 T 检验的一种推广。

和 5.3 节中一样，不论是大样本还是小样本，只要满足相应的 T 检验的条件，都可以应用 T 检验来检验两个总体之间的区别。根据待检验的两个样本之间的关系，两样本的 T 检验分为独立样本的 T 检验和配对样本的 T 检验两种。

所谓两独立样本是指两个样本所来自的总体相互独立，两个独立样本各自接受相同的测量，研究者或分析者的主要目的是分析两个独立样本的均值是否有显著的统计差异。例如，比较女性和男性的身高，教育从业者和金融从业者的起始工资等，都是两独立样本的例子。而配对的 T 检验则应用于比较同一个总体的两次不同的测量，例如医学研究中，比较药物的疗效；市场调查中，比较受调查者的父亲和母亲的教育程度等。

应用两独立样本 T 检验的前提条件如下。

- 独立性：两样本所来自的总体互相独立。
- 正态性：样本来自的两个总体应服从正态分布。大样本情况下，T 检验对正态性具有稳健性，也就是说，在样本所来自的总体不满足正态性条件时，如果两个样本的分布形状相似，它们的样本量相差不是太大并且样本量较大，仍然可以应用 T 检验。
- 方差齐性：方差齐性是指待比较的两个样本的方差相同。许多学者的仿真结果表明，如果两个组的样本量大致相等，略微偏离了方差齐性对检验结果的精度影响不大。在 T 检验中，SPSS 提供了方差齐性的 Levene 检验，当方差齐性不满足时，会提供方差齐性校正后的 T 检验结果。

本书技术支持网站的数据文件 `creditpromo.sav` 记录了接受不同促销方案的用户信用卡消费数据，现在需要检验新的促销方法是否能促进信用卡的消费，以此决定是否继续推进这种新促销方式。用统计的语言讲，就是比较采用新促销方法的信用卡消费金额均值和标准促销方法的信用卡消费金额均值，看二者是否在统计上有显著的差异。

打开数据文件 `creditpromo.sav`，数据如图 5-13 所示。

数据文件有三个变量，ID 是客户的 ID，insert 记录了客户接受的促销邮件类型。dollars 记录了促销期间客户的消费金额。在 5.4.1 节对数据进行初步的描述性统计分析，检查是否满足 T 检验的条件，5.4.2 节将介绍如何应用独立样本的 T 检验来分析新促销方案是否能促进用户的消费。

	id	insert	dollars	变量
1	148	标准	2232.77	
2	572	新促销	1403.81	
3	973	标准	2327.09	
4	1096	标准	1280.03	
5	1541	新促销	1513.56	
6	1947	新促销	1729.63	
7	2001	新促销	1609.71	
8	2130	标准	1476.62	
9	2616	标准	1460.77	
10	2886	新促销	1854.49	
11	3340	标准	1495.29	
12	3400	新促销	1107.21	
13	3539	标准	1949.59	

图 5-13 数据视图

5.4.1 数据初探

本节先对两种促销方式的客户消费数据进行描述性统计分析，初步探索两种不同的促销邮件下的客户花费情况。

选择【分析】→【描述统计】→【探索】，出现“探索性分析”对话框，如图 5-14 所示。把 dollars 选入到“因变量列表(D)”中，把“insert”选入到“因子列表(F)”中。



图 5-14 “探索性分析”对话框

单击【绘制(T)】按钮，在“描述性(D)”部分选择“直方图”，并且勾选“带检验的正态图(O)”，如图 5-15 所示。

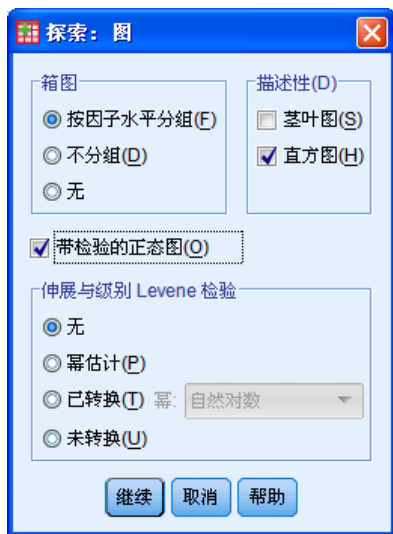


图 5-15 选择直方图和带检验的正态图

单击【继续】按钮，返回上级对话框，如图 5-14 所示。单击【确定】按钮。在结果输出窗口中，得到表 5-8、表 5-9、图 5-16 等输出。

以上操作过程可以通过以下的语法命令实现：

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\creditpromo.sav' .
DATASET NAME myData WINDOW=FRONT.
EXAMINE VARIABLES=dollars BY insert
  /PLOT BOXPLOT HISTOGRAM NPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

这里简单解释表 5-8 的“描述性统计分析报告”和表 5-9 的“正态性检验表”。

从描述性统计分析表可知，采用新促销方案的消费金额均值大于标准促销方案的消费金额均值。二者的标准差相差不大，二者标准差的比为 $346.67305/356.70317=0.972$ ，接近于 1。可以初步认为，二者的方差可能齐性，但是，两种促销方案对应的消费金额均值的差异是由于抽样的随机性造成的吗？如果不是，二者的确存在差别吗？要回答这些问题，必须借助统计检验。

表 5-8 描述性统计分析报告 (经过编辑)

接收到的邮件类型			统计量	标准误
促销期间的花费	标准	均值	1566.3890	21.92553
		均值的 95% 置信区 间	1523.2059 1609.5722	
		5% 修整均值	1565.5703	
		中值	1547.3610	
		标准差	346.67305	
		范围	2168.43	
		四分位距	435.66	
		偏度	.156	.154
		峰度	.566	.307
	新促销	均值	1637.5000	22.55989
		均值的 95% 置信区 间	1593.0674 1681.9325	
		5% 修整均值	1643.0651	
		中值	1661.0782	
		标准差	356.70317	
		范围	1892.23	
		四分位距	482.07	
		偏度	-.187	.154
		峰度	-.278	.307

图 5-16 分别是两种促销方案的直方图。从图 5-16 (A) 可以判断, 标准促销方案的分布是对称的, 应该为正态; 而 5-16 (B) 则略微偏离了对称性, 但是很难从直方图判断它是否服从正态性。而表 5-9 就是对两种邮件促销方式的正态性进行统计检验的结果。

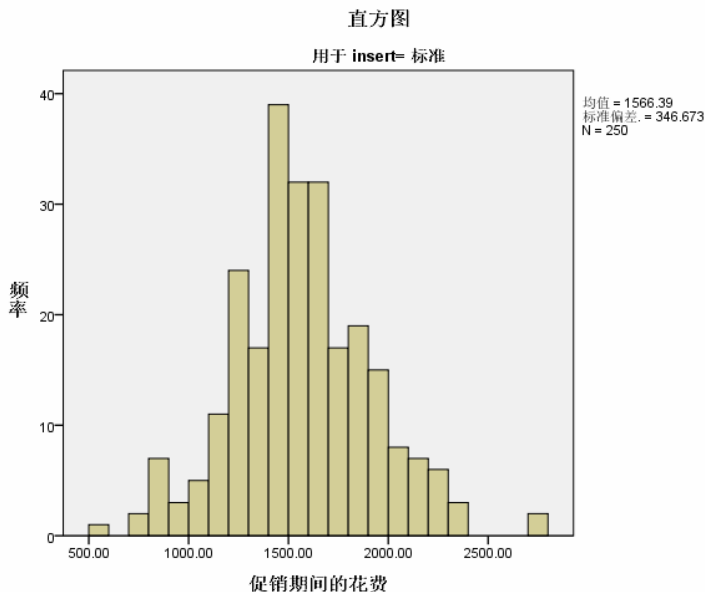


图 5-16 (A) 标准促销方案的花费直方图

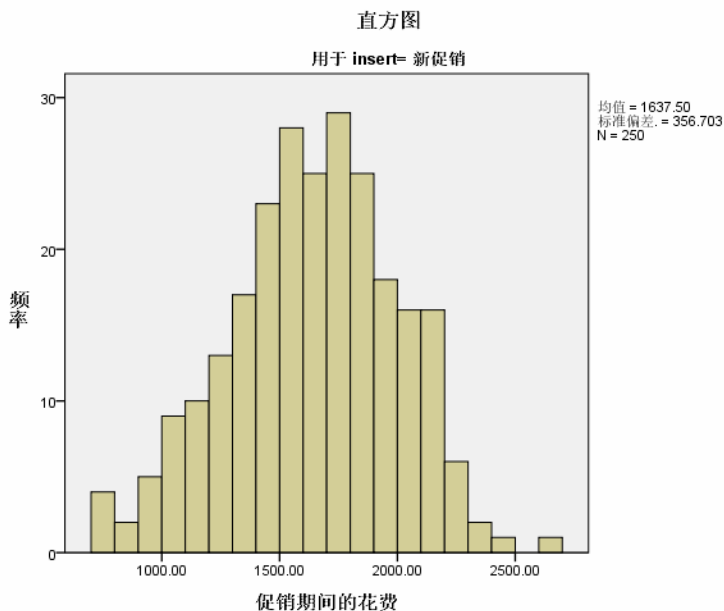


图 5-16 (B) 新促销方案的花费直方图

表 5-9 正态性检验表

接收到的邮件类型		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		统计量	df	Sig.	统计量	df	Sig.
促销期间的花费	标准	.054	250	.076	.992	250	.192
	新促销	.032	250	.200*	.992	250	.225

a. Lilliefors 显著水平修正

*. 这是真实显著水平的下限。

表 5-9 给出了两种正态性检验的结果。它们分别是带 Lilliefors 校正的 Kolmogorov-Smirnov 检验 (简称 K-S 检验) 和 Shapiro-Wilk 检验。对于这两个检验, 标准促销方案和新促销方案的 p 值都大于 0.05, 因此不能拒绝这两组样本分布的正态性假设。

基于描述性统计分析的结果, 标准促销方案和新促销方案都服从正态分布, 样本量都是 250, 二者的方差差别不大, 并且新促销和标准促销的样本是随机独立抽取的, 满足应用 T 检验的条件。因此我们可以应用独立样本的 T 检验来比较这两种促销方案是否有显著的区别。

5.4.2 T 检验

选择【分析】→【比较均值】→【独立样本 T 检验】, 进入独立样本 T 检验菜单和独立样本 T 检验的对话框, 如图 5-17 和图 5-18 所示。

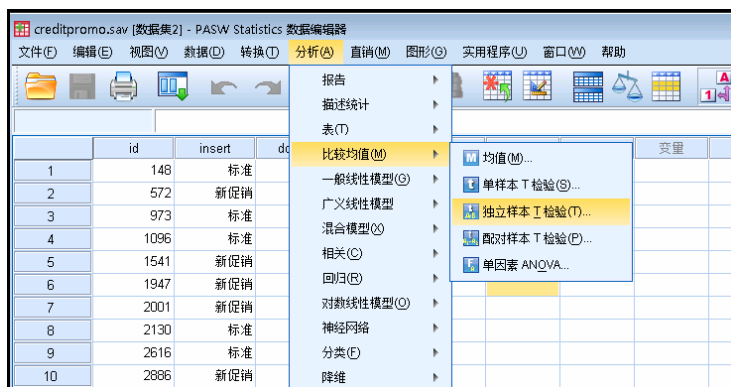


图 5-17 “独立样本的 T 检验”菜单



图 5-18 “独立样本 T 检验”对话框

把“促销期间的花费”选入“检验变量(T)”对话框，把“insert”促销方法变量选入“分组变量(G)”对话框。在“分组变量(G)”中的“Insert(?)”表示需要比较 insert 变量定义的分组，但是需要定义要比较哪两个组别。单击【**定义组(D)**】按钮，出现如图 5-19 所示的定义组对话框。在“使用指定值”部分输入两种促销活动对应的编码，这里在“组 1(1)”部分输入“0”，在“组 2(2)”部分输入“1”，如图 5-19 所示。

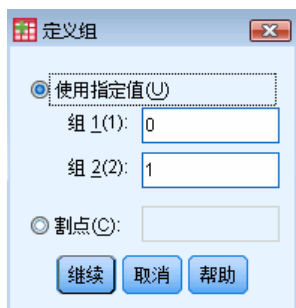


图 5-19 定义组

在图 5-19 中，也可以选择“割点(C)”，输入割点值，比如以 40 岁为割点，

比较 40 岁以下和 40 岁以上两组人群信用卡消费金额是否有显著不同。

单击【继续】按钮回到上级对话框，即如图 5-18 所示的窗口，单击【确定】按钮，查看输出结果，如表 5-10 所示。

表 5-10 独立样本 T 检验的输出结果

组统计量					
接收到的邮件类型		N	均值	标准差	均值的标准误
促销期间的花费	标准	250	1566.3890	346.67305	21.92553
	新促销	250	1637.5000	356.70317	22.55989

独立样本检验					
		促销期间的花费			
		假设方差相等		假设方差不相等	
方差方程的 Levene 检验	F	1.190			
	Sig.	.276			
均值方程的 t 检验	t	-2.260		-2.260	
	df	498		497.595	
	Sig.(双侧)	.024		.024	
	均值差值	-71.11095		-71.11095	
	标准误差值	31.45914		31.45914	
	差分的 95% 置信区间	下限		-132.91995	
		上限		-9.30183	

以上的操作过程也可以通过下列语法命令来实现。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\creditpromo.sav' .
DATASET NAME myData WINDOW=FRONT.
T-TEST GROUPS=insert(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=dollars
  /CRITERIA=CI(.95).
EXECUTE.
```

在“组统计量”表中显示两种促销方法信用卡花费的均值、标准差及均值的标准误。均值的标准误即为标准差除以样本量 N 的平方根。这里，标准促销方法的信用卡消费的标准差为 346.67305，而 $N = 250$ ，所以均值的标准误为 $346.67305 / \sqrt{250} = 21.92553$ 。

在“独立样本检验”表中显示方差齐性检验的结果和在方差齐性检验的不同结果下的 T 检验结果。“方差方程的 Levene 检验”的 p 值为 0.276，大于 0.1，说明两个独立样本的方差是齐性的。因此，我们选择“假设方差相等”列来查看假设检验的结果，该列的两独立样本 T 检验的“Sig.（双侧）”值为 0.024，小于显著性水

平 0.05, 说明新促销方法的消费金额显著不同于标准促销方法的消费金额。再结合不同促销方法的均值, 标准促销方法的消费均值为 1566.389, 而新促销方法的消费金额均值为 1637.5。因此相对于标准促销方法, 新促销方法确实能带来信用卡消费额的提高, 说明新促销方法在统计上确实比标准促销方法有效。

5.4.3 均值差的绘图

统计图也是比较两个总体均值差异的一种常用的方法。可以通过绘制箱图、误差图等来显示进行 T 检验的两个总体的均值之间的差异。首先绘制两个样本的箱图, 两种不同的促销方式的箱图如图 5-20 所示。从该箱图看出, 标准促销方式的箱图有 3 个离群值, 分别为记录 217、478 和 280, 它的箱体较新促销方式较短, 即后者的波动程度较大。新促销方式的箱体下边界较标准促销方式的箱图的下边界要高。

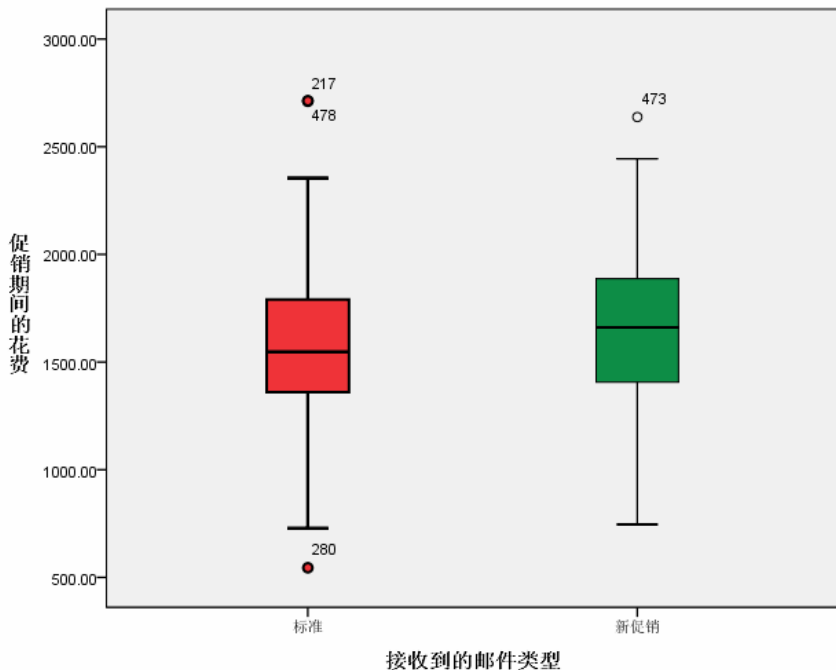


图 5-20 两个样本的箱图

可以通过下列代码来绘制以上图形。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\creditpromo.sav' .
DATASET NAME myData WINDOW=FRONT.
EXAMINE VARIABLES=dollars BY insert
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```

误差图绘制两种促销方式的均值以及均值的 95% 置信区间（默认为均值的 95% 置信区间，可以通过选项改变区间的置信水平）。从如图 5-21 所示的两个样本的误差图看出，新促销的误差图也较标准促销高，即新促销方式的总体均值要比标准促销方式要高。

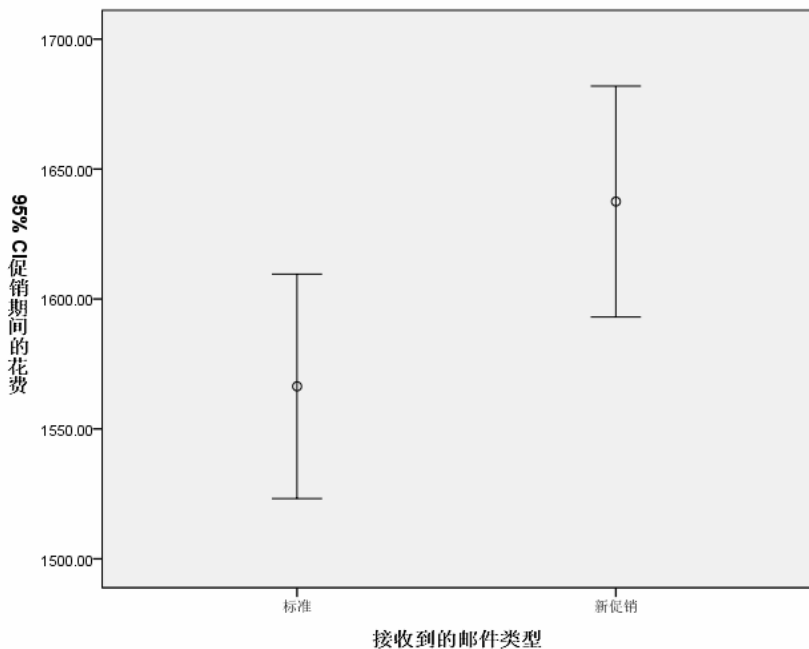


图 5-21 两个样本的误差图

可以通过下列代码来绘制以上误差图。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\creditpromo.sav' .
DATASET NAME myData WINDOW=FRONT.
GRAPH
  /ERRORBAR(CI 95)=dollars BY insert.
```

注意：许多情况下，用统计图来进行数据描述，可以增加对数据的理解。许多实际工作者建议，总是先把你的数据绘制成图表。

5.5 配对样本 T 检验

两配对样本 T 检验用来检验来自两配对总体的均值是否在统计上有显著性差异。在很多实验研究中，经常采用配对样本检验来检验某种配方或者试验手段的效果，常见的配对设计方法有以下几种：

- 同一受试对象处理前后的数据，例如服用某种药物前和服用之后的血压变化；
- 同一受试对象两个部位的数据，某种化妆品在一个人脸部不同位置的作用；
- 同一样本用两种方法测量的数据；
- 配对的两个受试对象分别接受两种处理后的数据。

应用两配对样本 T 检验的前提条件为：

- 两样本应是配对的。即受试对象的年龄、性别、体重、病况等非处理因素都相同或相似；
- 两个样本所来自的总体应服从正态分布（大样本情况下，T 检验对正态分布较为稳健）。

本章的数据文件 dietstudy.sav 包含对“Stillman diet”的研究结果。医生为检验某种饮食方案是否对有家庭心脏病史的病人有效，对 16 个病人进行了试验，记录他们在实行饮食方案前后的体重（磅）以及甘油三酸酯的水平（mg/100ml）。数据文件中的每个个案对应一个单独的病人。

现采用配对样本 T 检验对该饮食方案的效果进行分析。

打开数据文件 dietstudy.sav，然后选择【分析】→【比较均值】→【配对样本 T 检验（P）】，得到“配对样本 T 检验”对话框。如图 5-22 和图 5-23 所示。

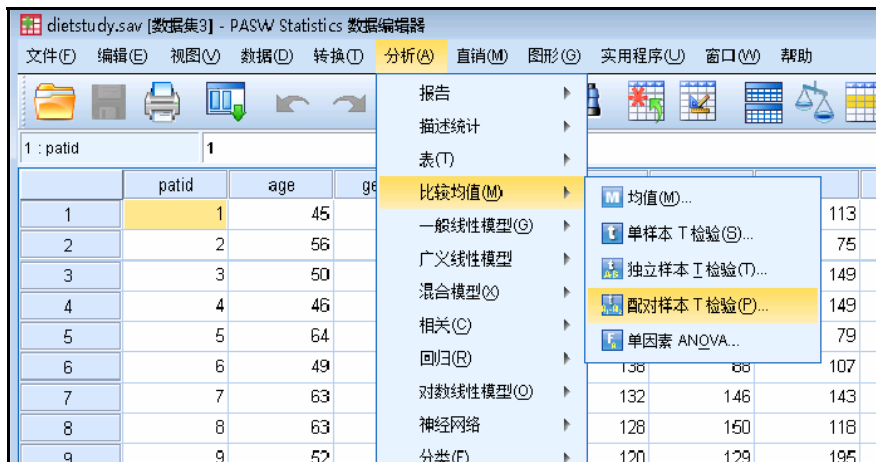


图 5-22 选择配对 T 检验



图 5-23 “配对 T 检验”对话框

在图 5-23 中，左边框中显示数据文件中的变量，右边框中显示配对的变量。在左边框中，同时选中“tg0”和“tg4”，然后单击指向右边的箭头，则在右边的“成对变量（V）”框中将显示该对变量。同样，把“wgt0”和“wgt4”作为另一对变量进行配对 T 检验。单击“确定”按钮，在输出窗口中查看分析结果，如表 5-11 所示。

表 5-11 配对 T 检验输出结果

成对样本统计量

	均值	N	标准差	均值的标准误
对 1 甘油三酸酯	138.44	16	29.040	7.260
最后的甘油三酸酯	124.38	16	29.412	7.353
对 2 体重	198.38	16	33.472	8.368
最后体重	190.31	16	33.508	8.377

成对样本相关系数

	N	相关系数	Sig.
对 1 甘油三酸酯 & 最后的甘油三酸酯	16	-.286	.283
对 2 体重 & 最后体重	16	.996	.000

成对样本检验

	成对差分					t	df	Sig.(双侧)
	均值	标准差	均值的标准误	差分的 95% 置信区间				
				下限	上限			
对 1 甘油三酸酯 - 最后的甘油三酸酯	14.063	46.875	11.719	-10.915	39.040	1.200	15	.249
对 2 体重 - 最后体重	8.063	2.886	.722	6.525	9.600	11.175	15	.000

以上的操作也可以通过下列语法命令完成：

```

NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 5\dietstudy.sav'
DATASET NAME myData WINDOW=FRONT.
T-TEST PAIRS=tg0 wgt0 WITH tg4 wgt4 (PAIRED)
/CRITERIA=CI(.9500)
/MISSING=ANALYSIS.
EXECUTE.

```

表 5-11 中，对每个分析变量，“成对样本统计量”表中输出两个配对样本的均值、样本量、标准差及均值的标准误。

“成对样本相关系数”表中输出两个配对样本的样本量、相关系数和相关系数的 p 值。“对 2”（即采用该饮食计划前的体重和采用该饮食计划后的体重）的相关系数为 0.996，从显著性值小于 0.05 可知，该相关系数明显大于 0，并且采用该饮食计划前体重和最后体重之间具有强线性相关，而采用该饮食计划前的甘油三酸酯含量和最后的甘油三酸酯含量的相关系数-0.286 的显著性值为 0.283，该相关系数不显著。

“成对样本检验”表输出配对样本的差值的均值、差值的标准差、差值均值的标准误、 t 统计量和相应的显著性值。这里， t 统计量的值为差值的均值除以均值的标准误。“对 1”的均值 14.063，即采用饮食计划前体重减去采用该饮食计划之后的体重的差值的均值，它表明采用该计划后的个体的甘油三酸酯含量减轻了 14.063。而体重减轻了 8.063，但由于甘油三酸酯的标准差及均值的标准误远远高于体重的标准差和标准误，因此“对 2”的 t 值远远大于“对 1”的 t 值。从最后的显著性值来看，体重的减轻在统计学上是显著的，即采用该饮食计划的受试者的体重确实减轻了，而受试者的甘油三酸酯在统计学上并没有显著变化。因此，对该饮食计划最终的评估结果为该减肥药可以减轻体重，但尚不能确定可以减轻甘油三酸酯（脂肪）。

应用配对 T 检验前，须注意下列事项：

需要先检查两个样本是否服从正态分布。等价地，可以检验两个配对样本的差值变量是否服从正态分布。可以应用直方图、Q-Q 图或者 K-S 检验等方法来检验差值变量的正态性。

要特别注意所分析变量中是否含有离群值。可以用箱图来检查离群值的情况。

有时候，可以先计算配对样本的差值变量，然后进行单样本的 T 检验。

5.6 小结

本章主要介绍均值过程，它给出变量的描述性统计量，同时可以给出相应总体的均值是否相等的判断。本章的重点内容为比较各组总体之间是否有统计上的差异，它包含单样本的 T 检验，独立样本的 T 检验和配对样本的 T 检验。在应用 T 检验进行均值比较之前，需要进行数据的初步分析，判断应用 T 检验的前提条件是否满足。另外，需要对 T 检验的结果进行详细的分析，得出合理的结论。

思考与练习

1. 数据文件 GSS2004_Mod.sav 中记录了男性或者女性每周上网浏览网页的时间（变量 WWWHR，单位小时）和每天观看电视的时间（变量 TVHOURS，单位小时）。用本章学习的技巧分析男性和女性在观看电视的时间和上网的时间上分别有什么区别。
2. 数据文件 GSS2004_Mod.sav 中记录了受访者的父亲和母亲的受教育的情况。试比较父亲的受教育情况（PAEDUC）和母亲的受教育情况（MAEDUC）是否不同，并给出父亲和母亲受教育年限的误差图。
3. 数据文件 GSS2004_Mod.sav 中记录了受访者的子女数（变量 CHILDS）、每周用于收发电子邮件的时间（变量 EMAILHR）和年龄（变量 AGE）。试对男性和女性的 CHILDS、EMAILHR 和 AGE 进行探索性分析，
 - （1）找出呈正态分布的变量。
 - （2）画出三个变量在不同性别的箱图，然后指出男性和女性的箱图的异同。
4. 配对两小样本做均值检验，应该使用 SPSS 哪种功能（假设两小样本都为正态分布）：
 - A) Means
 - B) One-Sample T Test
 - C) Independent-Samples T Test
 - D) Paired-Samples T Test
5. 对某一来自正态总体的样本与已知的总体均值比较，应该采取以下哪种 SPSS 分析方法：
 - A) Means
 - B) One-Sample T Test
 - C) Independent-Samples T Test
 - D) Paired-Samples T Test
6. 有关独立的 T 检验的论断，正确的是：
 - A) 在应用 Independent-Samples T Test 前，首先应该进行方差的齐性检验。如果方差齐性不满足，则无法应用该检验
 - B) 在应用 Independent-Samples T Test 前，首先应该进行方差的齐性检验。如果方差齐性不满足，则应该选用非参数检验

- C) 在应用 Independent-Samples T Test 前, 首先应该进行方差的齐性检验。
根据方差齐性检验的结果, 对分析结果进行相应的解读
- D) 以上都正确

参考文献

1. 梁冯珍, 关静等译. 统计学 (第 5 版). 北京: 机械工业出版社, 2009.
2. David S. Moore, George P. McCabe, Introductory to the Practice of Statistics.
3. Michael Sullivan, III, Statistics, informed decisions Using Data, Prentice Hall, 2004.
4. Rickman, R., N. Mitchell, J. Dingman, 和 J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. Journal of the American Medical Association, 228:54-58.

非参数检验

本章学习目标：

- 掌握应用各种非参数检验的条件；
- 掌握单样本非参数检验的各种方法及其区别；
- 掌握独立样本非参数检验的方法，能正确解释结果；
- 掌握相关样本非参数检验的方法，能正确解释结果。

6.1 非参数检验简介

参数检验方法检验的内容是总体分布的某些参数，例如均值、方差、比率等。当总体的分布已知时，统计上用参数检验的方法来检验两个总体的均值是否相等。从第 5 章可知，总体为正态分布时，可以应用 T 检验来检验均值之间的差异。T 检验本身是一种参数检验的方法。实际应用中，许多总体的分布不服从正态分布，亦不能通过变量转换而转化为正态分布，因此，不能够应用 T 检验来比较总体的均值。更进一步，在很多情况下，由于缺乏足够信息，总体的分布未知或难以确定。这些情况下，就不能使用参数检验的方法来进行分析。此时，应转而寻求更多的纯粹来自样本数据的信息，使用非参数检验方法。

非参数检验主要用于不考虑被研究对象的总体分布，或对总体的分布不做任何事先的假定，非参数检验的内容不是总体分布的某些参数，而是检验总体某些有关的性质，例如总体的分布位置、分布形状之间的比较，或者各样本所在总体是否独立等，因此该类统计方法被称为非参数检验。以均值比较为例，参数检验比较的是各样本的均值是否相等，而非参数检验比较的是各样本的中位数（中位数是分布位置的一种衡量）是否相等。

和参数检验方法相比，非参数检验方法具有以下优点：

- 稳健性：因对总体分布的约束条件放宽，从而对一些离群值或极端值不至于太敏感；
- 使用范围广：对数据的度量标准（或测量测度）无约束，定序数据、定量数据都可；部分数据缺失也可；小样本、分布未知样本、数据污染样本、混杂样本等都可以应用非参数方法。

但是非参数检验具有一个较大的缺点即检验效能较差，在参数检验中有显著性统计差异而在非参数检验中有可能并无显著性统计差异，因而在做分析时，应首先考虑参数检验，在无法满足参数检验方法的前提下，再考虑使用非参数检验方法。一般而言，非参数方法应用于以下场合：

- 参数检验方法的条件不满足。例如样本来自的总体不服从正态分布，T 检验不适用，必须应用非参数方法来比较两个总体的中心趋势；
- 研究定类变量和定序变量之间的关系。由于定类或者定序变量都不具有完备的运算性能，因此无法对总体分布作出假定或者检验总体的某种参数。

从版本 18 开始，SPSS 的非参数检验方法的用户界面发生了很大的变化。以前版本的非参数检验有 8 个子菜单，新的图形用户界面只有 4 个子菜单，它同时保留了以前版本的菜单以供老用户使用。更重要的是，非参数检验的输出结果界面采用了 SPSS 数据挖掘产品 Modeler 的输出风格，输出结果在模型浏览器中呈现，界面美观简洁，易于理解。SPSS 的非参数检验的用户界面，如图 6-1 所示。

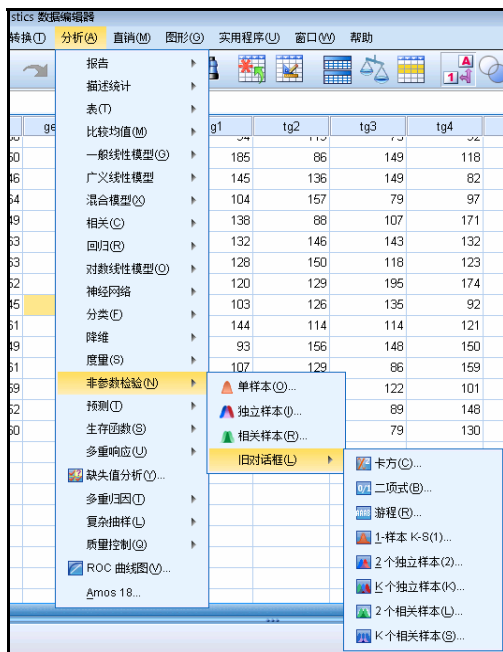


图 6-1 非参数检验用户界面

SPSS 版本 18 中非参数检验的用户界面和版本 17 以及以前版本不同,它把以前的卡方、二项式、游程、1-样本、2 个独立样本、2 个相关样本、K 个独立样本、K 个相关样本 8 个子菜单总括为单样本、独立样本、相关样本三个子菜单,这三个子菜单实现所有非参数检验的功能。在新的用户界面下,即使用户对非参数方法不甚了解,SPSS 能够智能地选择适用于所分析的数据的非参数方法。

注意:

- 新的用户界面统一了方法的选择,根据样本的个数来组织方法,简洁明了,输出结果用模型浏览器来展现,直观明了。它不能够选择输出描述性统计量和四分位数。
- 非参数统计过程仍然保留了 SPSS18 以前的非参数检验的界面,称为“旧对话框”,它的输出仍然为传统的表格方式展现检验结果,同时可以选择输出描述性统计量和四分位数。
- 在非参数检验过程的对话框和帮助文档中,把我们以前熟悉的变量 (Variable) 称为字段 (field)。

6.2 单样本非参数检验

得到一批数据之后,往往希望了解样本来自的总体的分布是否与某个已知的理论分布相吻合。这可以通过绘制样本数据的直方图、P-P 图、Q-Q 图等方法做粗略判断,还可以应用非参数检验的方法来实现。单样本非参数检验使用一个或多个非参数检验方法来识别单个总体的分布情况。非参数检验不需要待检验的数据呈正态分布。SPSS 的单样本非参数检验方法包括二项 (分布) 检验、卡方检验、Kolmogorov-Smirnov 检验 (以下简称 K-S 检验)、Wilcoxon 符号检验和游程检验五种非参数检验方法。选择【分析】→【非参数检验】→【单样本】,如图 6-1、图 6-2 所示。图 6-2 所示的单样本非参数检验的对话框有三个选项卡,它们分别为目标、字段和设置,其相应的功能如下。

- “目标”选项卡用于指定不同的检验目标设置。
- “字段”选项卡上指定字段分配。
- “设置”选项卡上指定专家设置。

1. 目标选项卡

在图 6-2 的“目标”选项卡,它询问“您的目标是什么”,它将设置非参数检验的目标。

- 自动比较观察数据和假设数据。SPSS 会根据“字段”选项卡中相应字段选择与其相适应的所有非参数检验方法，而在“设置”选项卡中也会自动选择“根据数据自动选择检验”。一般而言，如果待检验的变量是具有两个类别的分类变量，它将应用二项式检验比较观察数据和假设数据（原假设成立时应该出现的数据），对所有其他分类字段应用卡方检验，对连续字段应用 K-S 检验。该项为默认设置。
- 检验随机序列。该目标使用游程检验来检验观察到的随机数据值序列，用于判断观察到的数据值的随机性（或者判断观测值是否为白噪声）。选中该项后，在“设置”选项卡中相应的“游程检验”将被选择（被勾选）。
- 自定义分析。选中此项时，可以手动修改“设置”选项卡上的检验选项。注意，如果您随后在“设置”选项卡上更改了与当前选定目标不一致的选项，则“目标”选项卡中将会自动选择该设置。

另外，图 6-2 下方的“描述”部分是对选择的“目标”的简单的介绍。



图 6-2 单样本非参数检验

注意：根据设定的目标 SPSS 将自动选择相应的非参数分析方法，分析者无需选择特定的非参数检验方法，这大大节省了分析者的时间。

2. 字段选项卡

字段选项卡用于设置待检验的字段。

- 选中“使用预定义角色”时，则在 SPSS 数据文件的“变量视图”窗口中定义为“输入”、“目标”或者“两者都”角色的变量将自动进入“检验字段（T）”框中。定义为其他角色的变量将不会自动进入“检验字段”，

例如角色为“无”、“分区”或者“拆分”的变量将不会自动进入“检验”字段框中。

变量角色定义如图 6-3 所示，insert 变量的角色为“输入”。那么在单变量的非参数检验的“字段”选项卡中，insert 将自动进入“检验字段（T）”框中。如图 6-4 所示。

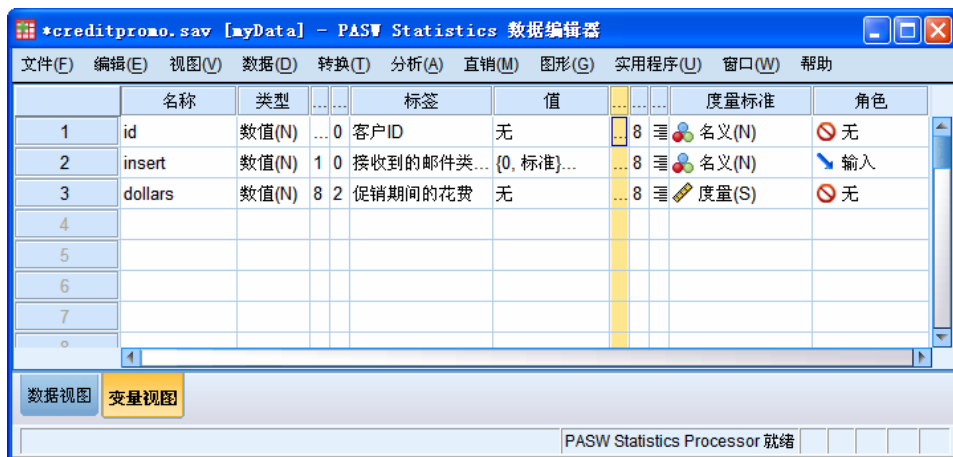


图 6-3 变量角色



图 6-4 字段选项卡-使用预定义角色

- 选中“使用定制字段分配”时，则可以手工选择进入“检验字段（T）”框中的变量。如图 6-5 所示。

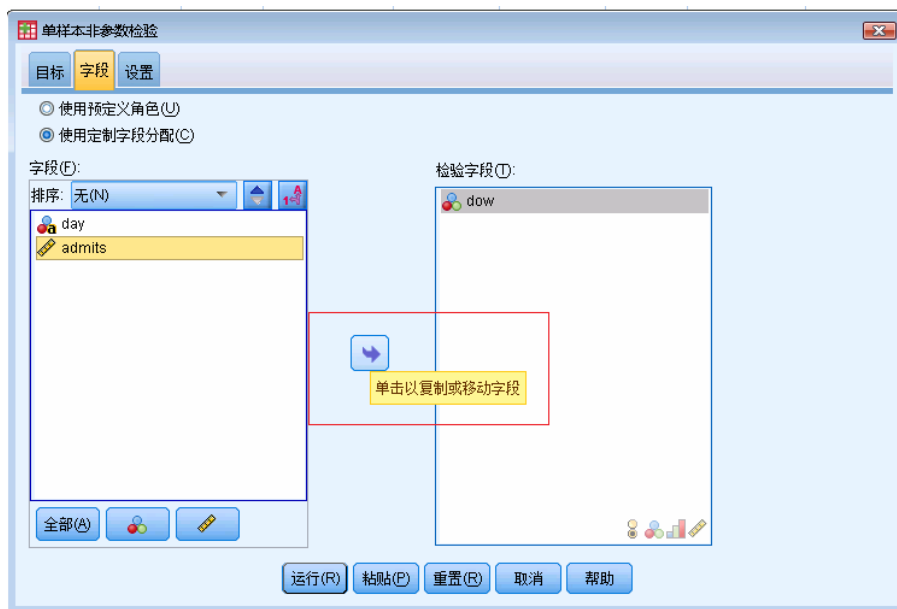


图 6-5 手工选择检验字段（即需要分析的变量）

注意：变量角色是在 SPSS 的数据挖掘产品 Clementine（现名 Modeler）中的一个概念，它定义变量在数据挖掘中所扮演的角色。现在，“角色”概念引入到 SPSS Statistics 中，把它作为变量的一个属性，它可以通过 SPSS 数据编辑器的“变量视图”窗口来更改。

可以选择的角色有：

- 输入：具有该角色的变量将作为分析或者建模的自变量；
- 目标：具有该角色的变量将作为分析或者建模的因变量或者目标变量；
- 两者都：具有该角色的变量既是自变量，也是因变量；
- 无：具有该角色的变量在分析中不起作用；
- 分区：具有该角色的变量将把数据集划分为训练集、检验集和测试集，默认训练集含数据集 70% 的个案，检验集含 30% 的个案；
- 拆分：该角色纯粹为和 SPSS Modeler 相互兼容。具有此角色的变量不会在 SPSS Statistics 中用作拆分文件变量。

3. 设置选项卡

“设置”选项卡用于指定要在所指定的检验字段（即待检验变量）上执行的检验，如图 6-6 所示。

- 根据数据自动选择检验。该设置对仅具有两个有效（非缺失）类别的分类字段应用二项式检验，对所有其他分类字段应用卡方检验，对连续字段应用 K-S 检验。它和“目标”选项卡中的第一个目标对应。
- 自定义检验（T）。该设置允许您选择要执行的特定检验。



图 6-6 设置选项卡

(1) 比较观察二分类可能性和假设二分类可能性（二项式检验）。二项式检验可以应用到分类变量，也可以应用于连续变量。可以检验标记字段（只有两个类别的分类字段）的观察分布是否与指定的二项式分布期望相同。也可以提供分类规则，把连续变量变为二分类变量，从而进行二项式检验。此外，可以输出置信区间。

(2) 比较观察可能性和假设可能性（卡方检验）。卡方检验可以应用到名义变量和有序变量。它可以根据变量类别的观察频率和期望频率间的差异来计算卡方统计量。

(3) 检验观察分布和假设分布（Kolmogorov-Smirnov 检验，简称 K-S 检验）。K-S 检验应用到连续变量（即尺度变量）。用于检验变量的样本累积分布函数是否为均匀分布、正态分布、泊松分布或指数分布。

(4) 比较观察中位数和假设中位数（Wilcoxon 符号秩检验）。Wilcoxon 符号秩检验可以应用到连续字段（即尺度变量）。这将生成一个变量中位数值的双样本检验。

(5) 检验随机序列（游程检验）。游程检验可以应用到所有测量类型的变量（这里称为字段）。它用于检验变量的值序列是否为随机序列。

6.2.1 卡方检验

卡方检验是一种常用的对总体分布进行检验的非参数检验方法。医生研究心脏

病人猝死人数与日期的关系，检验现在的人口结构和十年前是否一样，血型是否和人的性格有关系，现代社会中受过高等教育、高中毕业、初中毕业、小学毕业和文盲的比例是否为 3: 6: 10: 2: 1 等问题都可以通过卡方检验来实现。

卡方检验的原假设是：

H_0 ：样本来自的总体的分布与假设的分布（又称为期望分布或者理论分布）无显著差异。

卡方检验基本思想的理论依据是，如果从一个随机变量 X 所在的总体中随机抽取若干个观察样本，这些观察样本落在 X 的 k 个互不相交的子集中的观测频数服从一个多项分布，这个多项分布当 k 趋于无穷时近似服从卡方分布。基于这一思想，对变量 X 的总体分布的检验可以从对各个观察频数的分析入手。

如果变量 X 有 k 个互不相交的子集，在 H_0 成立的条件下，变量值落在第 i 个子集的频数设为 E_i ；设实际观测到的第 i 个子集的频数为 O_i ，则有以下 Pearson 卡方统计量

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

卡方统计量服从自由度为 $k-1$ 的卡方分布。如果卡方值较大，则说明期望频数与观测频数分布差距较大，没有证据支持原假设；反之，卡方值较小，说明期望频数与观测频数比较接近，不能拒绝原假设的论断。

本章的数据文件 `dischargedata.sav` 记录了 Winnipeg 医院每天的病人流量。医院管理者需要了解是否一周中每天的病人流量是相同的。

打开数据文件 `dischargedata.sav`，数据视图如图 6-7 所示。

	day	discharg	admits
1	星期天	44	68
2	星期一	78	87
3	星期二	90	90
4	星期三	94	84
5	星期四	89	82
6	星期五	110	84
7	星期六	84	71

图 6-7 数据视图

首先对每天的病人流量进行加权处理，用变量“discharge”(人均流量)为权重变

量对数据进行加权，如图 6-8 所示。



图 6-8 加权个案

把“discharge”（日均病人流量）作为频率变量，以日均病人流量作为权重对一周的各天进行加权，即该周每天都带有日均流量的权重系数，单击“确定”按钮。

以上操作可以通过下列语法命令实现：

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 6\dischargedata.sav' .
DATASET NAME myData WINDOW=FRONT.
WEIGHT BY discharge.
```

进入单样本非参数检验菜单，如图 6-9 所示。

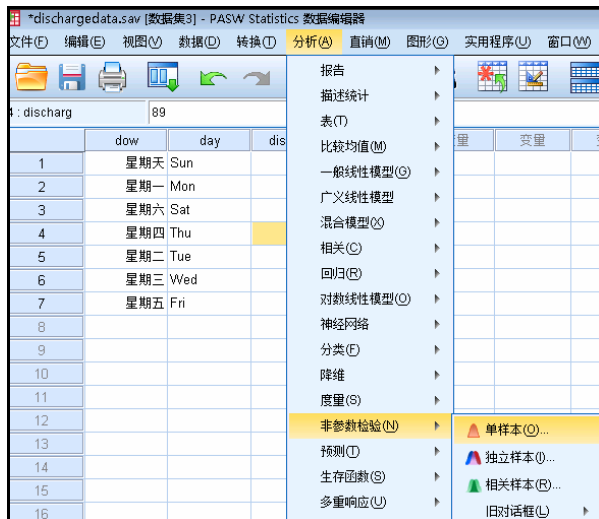


图 6-9 单样本非参数检验

“目标”选项卡保留默认设置，即“自动比较观察数据和假设数据（U）”，

如图 6-10 所示。



图 6-10 目标选项卡

“字段”选项卡中选择“使用定制字段分配”，并把“day”（星期几）选入“检验字段”，如图 6-11 所示。



图 6-11 字段选项卡

在“设置”选项卡中选择“自定义检验”，并选择“比较观察可能性和假设可能性（卡方检验）”，如图 6-12 所示。单击该项下的【选项】按钮，如图 6-13 所

示，可以设置卡方检验的假设可能性——即所检验类别的先验概率，或者理论概率。这里，保留默认值“所有类别的概率相等”。



图 6-12 设置选项卡

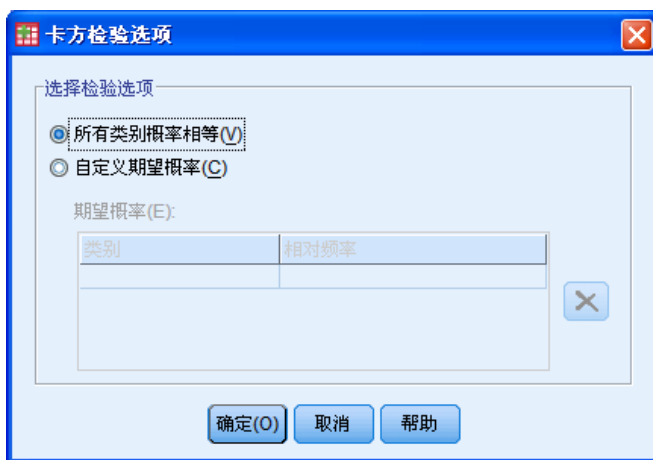


图 6-13 选择卡方检验的先验概率

在图 6-13 中，单击【确定】按钮，返回上级对话框（如图 6-12 所示），然后单击【运行】按钮。

以上操作可以通过下列语法命令来完成。

```

DATASET ACTIVATE myData.
*Nonparametric Tests: One Sample.
NPTESTS
  /ONESAMPLE TEST (day) CHISQUARE(EXPECTED=EQUAL)
  
```



```

/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95.
EXECUTE.

```

在输出浏览器窗口中得到结果如图 6-14 所示。

	原假设	检验	显著性水平	决策
1	星期几 的类别发生概率相等。	单样本卡方检验	.000	拒绝原假设。

显示渐近显著性。显著性水平为 .05。

图 6-14 卡方检验结果-假设检验摘要

从图 6-14 可见，假设检验摘要列出了检验的原假设、检验方法、显著性水平（即 P 值）和决策。该检验的 P 值为 0.000，小于 0.05，决策为“拒绝原假设”，说明该周各天的日均病人流量有显著差异，每天病人的流量是不等的。双击结果浏览器中的图 6-14，可以进入模型浏览器来观察更详细的分析结果，如图 6-15 所示。



图 6-15 模型浏览器

如图 6-15 所示的模型浏览器为互动图。可以在模型浏览器下面的选项框中选择输出的内容。从图 6-15 右侧下方的表格可直观地看到，检验的卡方统计量为 29.389， P 值为 0.000，通过了显著性统计检验，即每天的人流量有显著区别。其中星期五的人流量最多，星期天的人流量最少（一般医院星期天没有普通门诊，而大部分的人星期天也会待在家里休息），其他几天日均流量差别不是特别大。

注意：“卡方检验”的模型浏览器视图显示聚类条形图和检验表。

聚类条形图显示检验字段每个类别的观察频率和假设频率。悬停在条形上将显示在工具提示中显示观察频率和假设频率及其差别（残差）。观察和假设条形中的可见区别表明检验字段可能没有假设的分布。

6.2.2 二项式检验

现实生活中很多数据的取值是二值的，例如，性别变量有男性和女性两个取值；产品有合格和不合格两个取值；筛子可以有偶数面和奇数面两个取值。通常将二值分别用 0 和 1 表示。如果一个试验只有两个结果（分别称它们为失败和成功，并分别用 0 和 1 来表示），并且每次试验中每个结果出现的概率是固定的，则该试验为 0-1 试验（或称为贝努力试验）。如果将 0-1 试验独立地重复进行 n 次，则得到 n 重贝努力试验。在一个 n 重贝努力试验中，结果 1 出现的次数 X 是一个随机变量，它所服从的概率分布称为二项分布。

二项分布记为 $B(n, p)$ ，其中 n 为重复试验的次数， p 为一次试验中出现结果 1 的概率（或者成功的概率），其概率密度函数为：

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, 2, \dots, n \quad (\text{二项分布公式})$$

SPSS 的二项式检验通过样本数据检验样本来自的总体是否服从指定的二项分布。例如，现代社会男、女的比例是否为 1.01: 1；工厂的次品率是否为 1%等都可以通过二项式检验完成。

一家电信公司每个月大约有 27% 的用户会离开，为减少客户流失，公司经理想了解不同的客户群的流失比例是否有差异。客户流失数据在文件 `telco.sav` 中。我们所关心的是流失客户，即 `Churn` 值为 1 的客户。首先把个案按照客户类型和是否流失排序，这样每一类客户中的第一条个案即为流失客户。然后按照客户类型来分隔文件，最后用二项式检验各个客户群的流失比例是否有差异。

注意：SPSS 二项式检验首先需要定义“成功”和“失败”类别：

- 如果是分类变量，SPSS 二项式检验默认数据中的第一个类别为成功类别；
 - 如果是连续变量，则把小于或等于样本中点的值作为成功类别；
 - 在二项式检验的选项中可以更改默认的设置。
-

1. 数据排序

打开本章的数据文件 telco.sav，选择【数据】→【排序个案】，如图 6-16 所示。



图 6-16 排序个案

选择“custcat”和“churn”两个变量到排序依据框中，顺序如图 6-16 所示。选中排序依据框中的“churn”变量，在排列顺序框中选择“降序”。单击【确定】按钮。排序后的数据视图如图 6-17 所示。

	region	tenure	age	gender	custcat	churn
1	Zone 2	13	44	Male	Basic service	Yes
2	Zone 2	33	33	Female	Basic service	Yes
3	Zone 2	5	33	Female	Basic service	Yes
4	Zone 3	9	34	Female	Basic service	Yes
5	Zone 2	30	34	Male	Basic service	Yes
6	Zone 2	36	45	Female	Basic service	Yes
7	Zone 3	3	43	Female	Basic service	Yes
8	Zone 2	2	31	Male	Basic service	Yes
9	Zone 3	3	20	Female	Basic service	Yes
10	Zone 3	3	31	Male	Basic service	Yes
11	Zone 3	19	29	Male	Basic service	Yes
12	Zone 2	7	35	Male	Basic service	Yes
13	Zone 2	7	33	Female	Basic service	Yes
14	Zone 3	7	24	Male	Basic service	Yes
15	Zone 1	3	33	Female	Basic service	Yes
16	Zone 1	17	51	Female	Basic service	Yes
17	Zone 1	3	33	Male	Basic service	Yes
18	Zone 3	39	24	Male	Basic service	Yes
19	Zone 2	6	30	Female	Basic service	Yes
20	Zone 2	1	21	Female	Basic service	Yes

图 6-17 排序后数据

2. 分隔文件

我们按照客户类型来分隔数据文件。选择【数据】→【拆分文件】，如图 6-18 所示。选择“比较组 (C)”，把 custcat 变量选入到“分组方式”框中。由于在 1. 中已经对文件排序，因此我们选择“文件已排序”。设置完成后，单击【确定】按钮。



图 6-18 拆分文件

以上操作可以通过下列语法命令完成。

```

DATASET ACTIVATE myData.
SPLIT FILE LAYERED BY custcat.
EXECUTE.

```

3. 二分类变量二项式检验

选择【分析】→【非参数检验】→【单样本】，如图 6-19 和图 6-20 所示。在“字段”选项卡中，选择“churn”为检验字段，如图 6-19 所示。在“设置”选项卡中，勾选“自定义检验 (T)”，选中“比较观察二分类可能性和假设二分类可能性（二项式检验） (O)”，如图 6-20 所示。

在图 6-20 中，单击“比较观察二分类可能性和假设二分类可能性（二项式检验） (O)”下面的【选项 (B)】按钮，设置二项式检验的假设比例（或者概率），即二项分布公式中的概率 p ，如图 6-21 所示。在“假设比例”部分，输入待检验的比例—0.27，其他设置保留默认值。



图 6-19 设置检验字段



图 6-20 设置检验方法

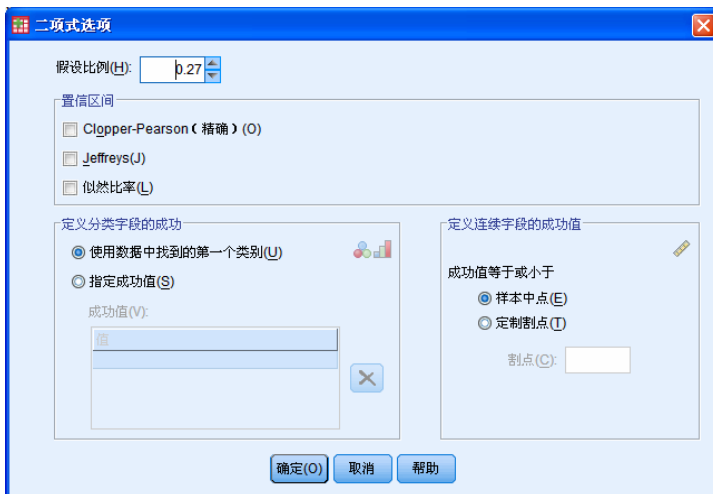


图 6-21 设置检验比例并定义成功事件

如图 6-21 设置后,单击【确定】按钮,返回上级对话框,如图 6-20 所示。单击左边“选择项目(S)”框中的“检验选项”,可以设置检验的显著性水平,即犯第一类错误的概率 α 和置信区间的置信水平。如图 6-22 所示,我们保留默认值。



图 6-22 设置检验选项-显著性水平和置信水平

单击设置“选项卡”左边的“选择项目(S)”框中的“用户缺失值”,可以设置分类变量缺失值的处理方法。如图 6-23 所示,如果选择“包括”,则缺失值将被作为一个类别,我们保留默认值。



图 6-23 缺失值处理方法

设置完毕后,单击【运行】按钮。输出浏览器中的输出和卡方检验的输出类似。

先给出假设检验摘要，如图 6-24 所示。双击假设检验摘要进入模型浏览器中，如图 6-25 和图 6-26 所示。

客户类型 = Basic service

假设检验摘要			
	原假设	检验	显著性水平 决策
1	由上个月流失与否 = Yes 和 No 定义的类别发生概率为 0.27 和 0.73。	单样本二项式检验	.070 保留原假设。

显示渐近显著性。显著性水平为 .05。

客户类型 = Plus service

假设检验摘要			
	原假设	检验	显著性水平 决策
1	由上个月流失与否 = Yes 和 No 定义的类别发生概率为 0.27 和 0.73。	单样本二项式检验	.000 拒绝原假设。

显示渐近显著性。显著性水平为 .05。

客户类型 = E-service

假设检验摘要			
	原假设	检验	显著性水平 决策
1	由上个月流失与否 = Yes 和 No 定义的类别发生概率为 0.27 和 0.73。	单样本二项式检验	.500 保留原假设。

显示渐近显著性。显著性水平为 .05。

客户类型 = Total service

假设检验摘要			
	原假设	检验	显著性水平 决策
1	由上个月流失与否 = Yes 和 No 定义的类别发生概率为 0.27 和 0.73。	单样本二项式检验	.000 拒绝原假设。

显示渐近显著性。显著性水平为 .05。

图 6-24 (A) 二项式检验摘要

图 6-24 (B) 二项式检验摘要

以上操作可以通过下列语法命令来完成。

```
DATASET ACTIVATE myData.
*Nonparametric Tests: One Sample.
NPTESTS
  /ONESAMPLE TEST (churn) BINOMIAL(TESTVALUE=0.27 SUCCESSCATEGORICAL=FIRST
    SUCCESSCONTINUOUS=CUTPOINT(MIDPOINT))
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05 CILEVEL=95.
```

从假设检验摘要可知，客户类型 Plus service 和 Total service 的显著性水平值小于 0.05，二项式检验显著，决策为“拒绝原假设”，即他们的客户流失比例显著不等于 27%。

分别双击客户类型 Basic service 和 Plus service 的假设检验摘要图，得到图 6-25 和图 6-26。

图 6-25 和图 6-26 的模型浏览器中的“二项式检验”显示堆积条形图和检验表。

注意：二项式检验的输出结果中的堆积条形图显示检验字段“成功”和“失败”类别的观察频率和假设频率，其中“失败”堆积在“成功”的顶部。鼠标滑轨条形时，悬停在条形旁的框中将提示该类别百分比。观察值条形和假设值条形中的明显区别表明检验字段可能没有假设的二项式分布。

图 6-25 展现了二项式检验的检验统计量、标准误、显著性水平等，Basic service 客户群的二项式检验的显著性水平为 0.070，大于 0.05，因此决策为“保留原假设”。直方图显示观察值和假设值的流失比例大致相等。

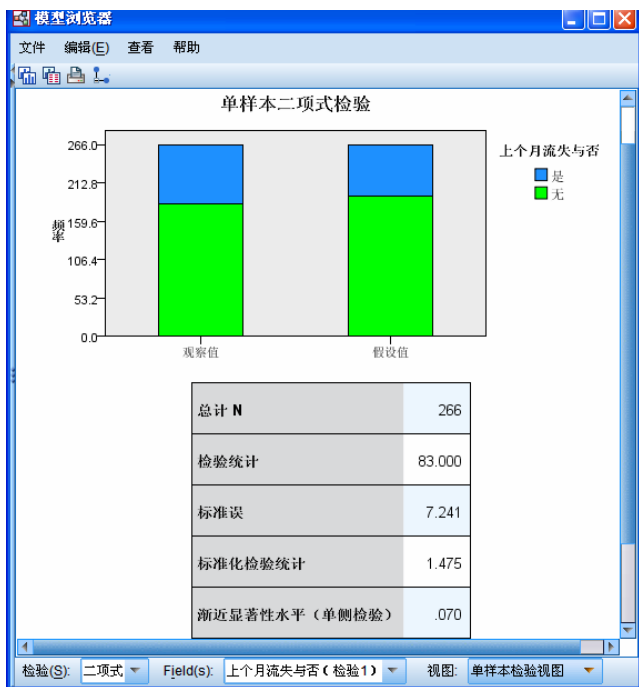


图 6-25 Basic service 客户群二项式检验结果



图 6-26 Plus service 客户群二项式检验结果

图 6-26 展示了客户群 Plus service 的二项式检验结果，单侧显著性水平为 0.000（SPSS 中小于 0.0005 的值显示为 0.000），因此拒绝原假设，即该客户群中流失的比例小于 27%。

从图 6-26 上方的堆积条形图可见，观察流失比例和假设流失比例相差很大。

注意：应用新的非参数检验的界面模型，在模型浏览器中通过图形来展示结果，形象直观，一目了然。如果需要了解更多的信息，比如知道 Total service 和 Plus service 两个客户群的流失比例不同于 27%，需要更进一步知道这两个组哪个流失严重，则需要用其他方法继续进行分析。

实际上，Plus service 的流失比例（为 16%）显著小于 27%。而 Total service 的流失比例（为 37%）显著大于 27%，该客户群是流失风险最大的，需要重点关注该组客户，找出它们不满意的原因，进而降低该组客户的流失。

4. 连续变量的二项式检验—旧对话框实现

在 SPSS 版本 18 以前的非参数检验，都可以在“旧对话框”菜单中找到，如图 6-27 所示。

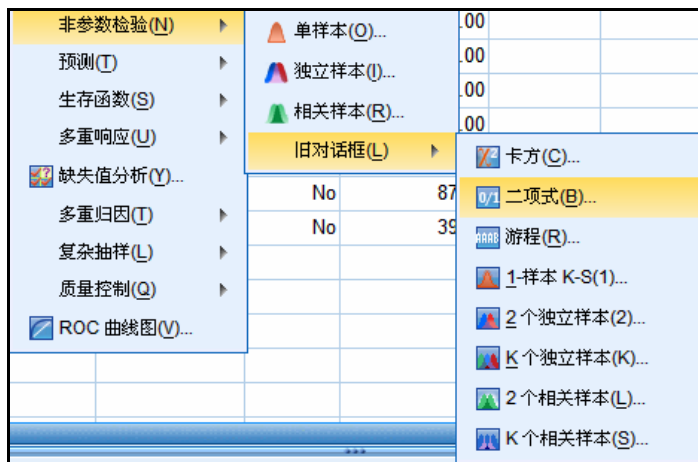


图 6-27 非参数检验旧对话框

在 3 中，我们知道各个客户群流失的比例不同，其中 Total service 组的流失比例显著大于 27%。我们想知道，收入的高低和流失是否有关系。在流失的客户和没有流失的客户中，收入在中位数\$47,000 以上的家庭和在\$47,000 以下的家庭所占的比例是否有显著差异呢？

我们先按照流失与否对数据进行分隔，如图 6-28 所示。“分组方式（G）”选择 churn，同时选中“按分组变量排序文件”，单击【确定】按钮。



图 6-28 分割文件

选择【非参数检验】→【旧对话框】→【二项式】，进入“二项式检验”对话框，如图 6-29 所示。把 income 变量选入到“检验变量列表(T)”框中，在左下部分的“定义二分法”中，选择“割点”，输入 47，即取 income 的中位数 47（代表 \$47,000）作为分割点。检验比例取默认值 0.5。意即比较收入小于或等于 \$47,000 的家庭和大于 \$47,000 的家庭的比例是否显著区别于 0.5，也就是说，检验在流失客户群中家庭收入是否和流失有关系。

在图 6-29 中，单击“选项”按钮，设置二项式检验输出的统计量和对缺失值处理方式。如图 6-30 所示，这里选择“描述性”和“四分位数”。



图 6-29 定义成功和失败事件-割点

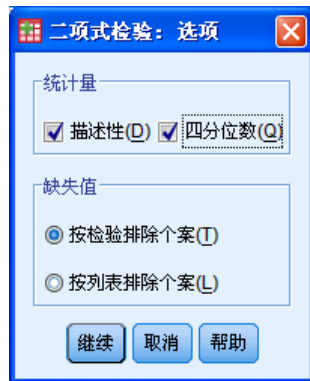


图 6-30 二项式检验选项

注意：新的用户界面没有该选项，因此不能输出描述性统计量和四分位数。

图 6-30 设置完毕后，单击【继续】按钮，返回上级对话框图 6-29，单击【确定】按钮。

以上操作可以通过以下语法命令完成。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 6\telco.sav' .
DATASET NAME myData WINDOW=FRONT.
SORT CASES BY churn.
SPLIT FILE LAYERED BY churn.
NPAR TESTS
  /BINOMIAL (0.50)=income (47)
  /STATISTICS DESCRIPTIVES QUANTILES
  /MISSING ANALYSIS.
```

二项式检验输出结果如表 6-1 所示。

表 6-1 二项式检验输出结果

描述性统计量						
		上个月流失与否				
		无	是			
		家庭收入（千美元）	家庭收入（千美元）			
N		726	274			
均值		83.5386	61.6277			
标准差		119.40447	60.97078			
极小值		9.00	9.00			
极大值		1668.00	429.00			
百分位	第 25 个	30.0000	26.7500			
	第 50 个（中值）	49.0000	41.0000			
	第 75 个	89.2500	70.0000			

二项式检验						
上个月流失与否		类别	N	观察比例	检验比例	渐近显著性(双侧)
无	家庭收入（千美元） 组 1	<= 47	345	.48	.50	.194 ^a
	组 2	> 47	381	.52		
	总数		726	1.00		
是	家庭收入（千美元） 组 1	<= 47	160	.58	.50	.006 ^a
	组 2	> 47	114	.42		
	总数		274	1.00		

a. 基于 Z 近似值。

从表 6-1 输出结果的“二项式检验”部分可知，没有客户流失的组中家庭收入高于中位数 47 和小于等于 47 的比例与 50% 无显著区别；而有客户流失的组中，小于等于 47 的客户显然超过大于 47 的客户，即收入偏低的客户居多。从“描述性统

计量”部分知，在流失组中其家庭收入均值为 61.63，没有流失的组均值为 83.54。流失组的三个四分位数都小于没有流失组的相应的四分位数。

6.2.3 K-S 检验

K-S 检验是一种利用样本数据推断样本来自的总体是否与某一理论分布有显著差异的非参数统计方法，是拟合优度检验的方法之一。它适用于探索连续型随机变量的分布。

K-S 检验在实际中有广泛的应用，例如可以检验某个班级某科的成绩是否与正态分布有显著差异，某地区新生婴儿的体重是否与正态分布有显著差异，某商店顾客的到来是否与泊松分布有显著差异等都可以用 K-S 检验来实现。SPSS 的 K-S 检验可以检验四种理论分布：正态分布、均匀分布、泊松分布和指数分布。

单样本 K-S 检验的原假设为：

H_0 ：样本来自的总体与指定的理论分布无显著差异。

精算师需要分析某个地区驾驶员的交通事故数量，她在该地区随机抽取了 500 名驾驶员的数据。她想验证驾驶员的交通事故数量是否服从泊松分布。我们采用单样本的 K-S 检验。

打开本章的数据文件 autoaccidents.sav，选择【分析】→【非参数检验】→【单样本非参数检验】。如图 6-31 所示，在“字段”选项卡中设置检验字段为“事故数”。



图 6-31 设置检验字段

在“设置”选项卡中设置检验方法为 K-S 检验，如图 6-32 所示。



图 6-32 设置检验方法

在图 6-32 中，单击“K-S”选项下面的【选项（K）】按钮，设置检验的分布类型，我们选择“泊松”，SPSS 默认分布的参数是从样本数据中计算而来，也可以手工输入该分布的参数，这里我们保留默认值，如图 6-33 所示。



图 6-33 设置待检验的分布类型

单击【确定】按钮，返回上级对话框（如图 6-32 所示），单击【运行】按钮。

以上操作可以通过下列语法命令来实现。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 6\autoaccidents.sav' .
DATASET NAME myData WINDOW=FRONT.
NPTESTS
  /ONESAMPLE TEST (事故数) KOLMOGOROV_SMIRNOV(POISSON=SAMPLE )
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05 CILEVEL=95.
```

在结果浏览器中，K-S 检验的结果如图 6-34 和图 6-35 所示。

假设检验摘要			
	原假设	检验	显著性水平 决策
1	过去5年的事故数的分布为具有均值 1.722 的泊松分布。	单样本 Kolmogorov-Smirnov 检验	.028 拒绝原假设。

显示渐近显著性。显著性水平为 .05。

图 6-34 K-S 检验摘要

显著性水平为 0.028，小于 0.05，因此决策为拒绝“过去 5 年的事故数的分布为具有均值 1.722 的泊松分布”的原假设。

双击如图 6-34 所示的 K-S 检验摘要，进入模型浏览器，如图 6-35 所示。

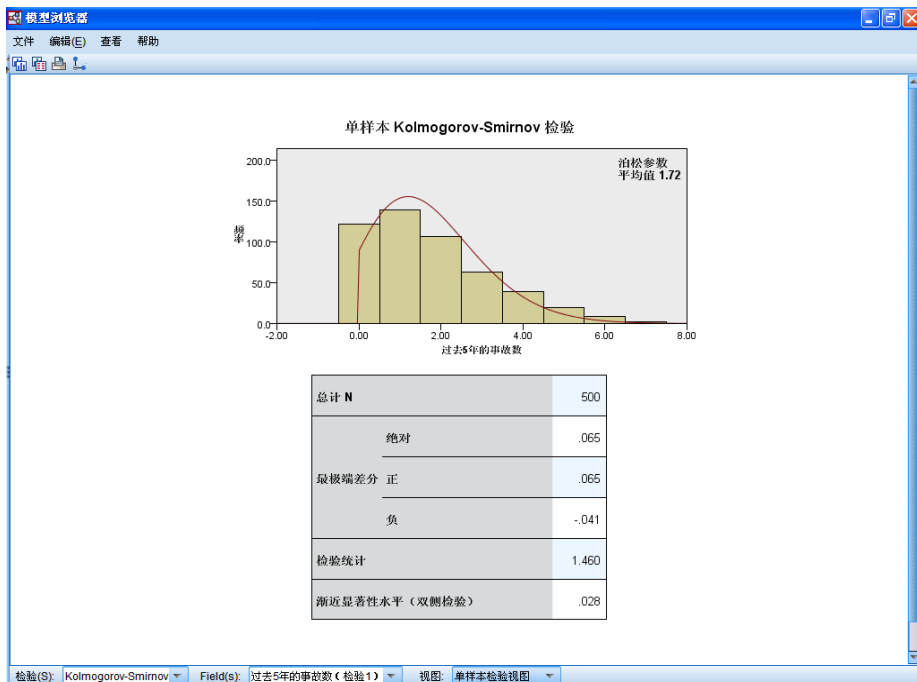


图 6-35 检验统计量

图 6-35 给出了 K-S 检验的检验统计量和检验变量事故数的直方图。

注意：K-S 检验模型浏览器视图显示直方图和检验表。直方图包括假设均匀、正态、泊松或指数分布概率密度函数的重叠。注意，该检验基于累积分布，同时表格中报告的“最极端差分”应相对于累积分布进行解释。

6.2.4 Wilcoxon 符号秩检验

Wilcoxon 符号秩检验用于检验样本所来自的总体的中位数和所给的值没有显著区别。该检验适用于连续型数据（或者尺度数据），它把观测值和原假设的中心位置之差的绝对值的秩分别按照不同的符号相加作为其检验统计量。注意，该检验需要假定样本数据来自分布连续对称的总体，此时总体中位数等于均值。

Wilcoxon 符号秩检验的原假设为：

H_0 ：样本所来自的总体的中位数等于给定的数值。

本章的数据文件 alcohol.sav 是欧洲城镇每人每年平均消费的酒类相当于纯酒精数。我们用 Wilcoxon 符号秩检验来考察人均年消费酒精量的中位数等于纯酒精 8 升。Wilcoxon 符号秩检验的操作方式和以上介绍的单样本的非参数检验十分类似，留给读者自己完成。我们对解读模型浏览器中的结果作简单介绍。

“Wilcoxon 符号秩次检验”模型浏览器视图显示直方图和检验表。直方图包括显示观察和假设中位数的垂直线。如果两条垂直线的距离过远，有理由怀疑观测中位数和假设中位数无显著差别这一原假设。Wilcoxon 符号秩次检验结果如图 6-36 所示。

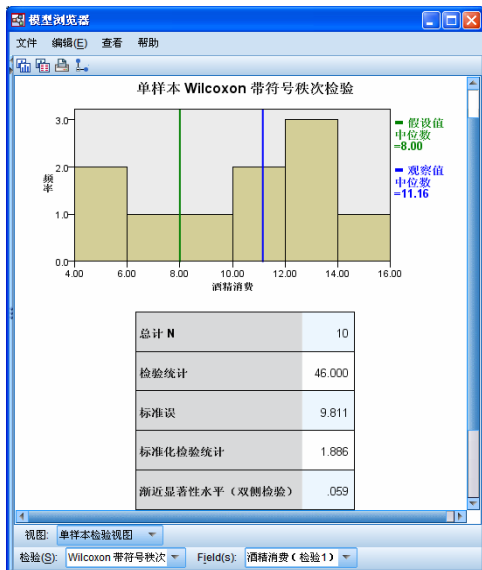


图 6-36 Wilcoxon 符号秩检验结果

6.2.5 游程检验

游程检验用于检验某一变量的两个值的出现顺序是否随机，对于连续型变量的随机性检验也可以转化为只有两个取值的分类变量的随机性的检验。游程检验通过对样本观测值的分析，用来检验该样本所来自的总体序列是否为随机序列（又称为白噪声序列）。它也可以用来检验一个样本的观测值之间是否相互独立。

游程检验的原假设为：

H_0 ：总体中变量值的出现是随机的。

假定我们投掷一枚硬币，以概率 p 得到正面，记为 1，以概率 $1-p$ 得到反面，记为 0。

例如，下面为投掷 23 次得到的结果：

00000001111110000111100

连在一起的 0 或者 1 称为一个游程。以上序列有 3 个 0 游程，2 个 1 游程，共有 5 个游程。如果该试验是随机的，则不太可能出现许多 1 或许多 0 连在一起，也不可能 0 和 1 交替出现得太频繁。游程数太大或者太小都将表明变量取值存在不随机现象。

SPSS 单样本非参数检验中的游程检验利用游程数构造检验统计量。如果 n_1 为出现 0 的个数， n_2 为出现 1 的个数， r 表示总的游程数。当 n_1 和 n_2 较大时，总的游程数的抽样分布的均值和方差分别为：

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} \text{ 和 } \sigma_r^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

则检验统计量为，

$$Z = \frac{r - \mu_r}{\sigma_r}$$

检验数据文件 runs.sav 中的 runs 变量的取值是否为随机的。游程检验的操作方式和以上介绍的单样本的非参数检验十分类似，留给读者自己完成。我们给出游程检验的结果及其解释，如图 6-37 和图 6-38 所示。

假设检验摘要

	原假设	检验	显著性水平	决策
1	由 runs ≤ 1 和 >1 定义的值序列为随机序列。	单样本游程检验	.003	拒绝原假设。

显示渐近显著性。显著性水平为 .05。

图 6-37 假设检验摘要

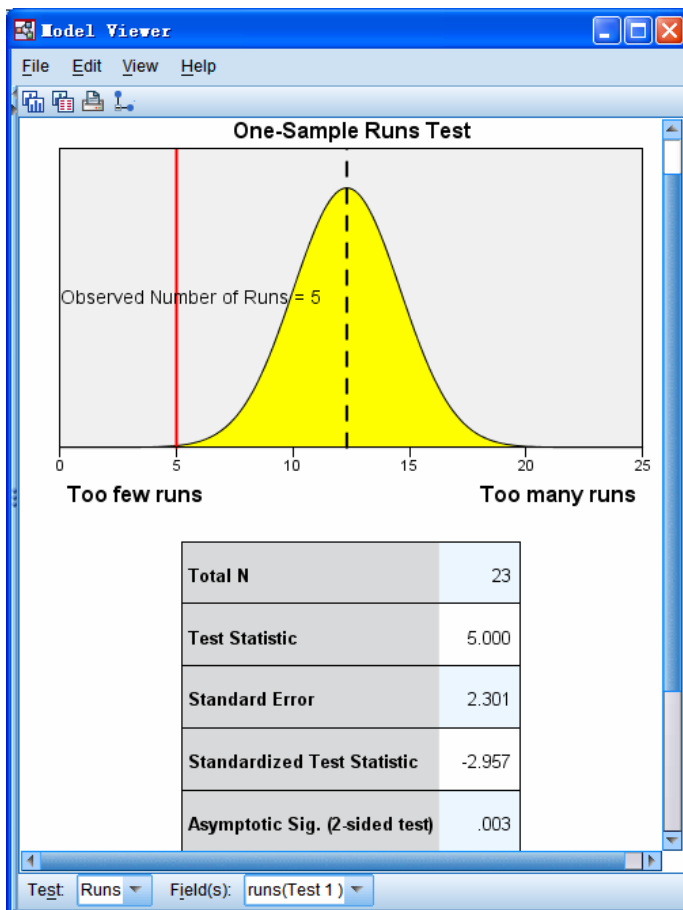


图 6-38 游程检验结果

注意：“游程检验”视图显示图表和检验表。显示以垂直线标记的观察到的游程数的正态分布。注意，当执行精确检验时，该检验不基于正态分布。

6.3 独立样本非参数检验

独立样本非参数检验使用一个或多个非参数检验方法来识别两个或更多个组间的差别。对于两个分布未知的总体，或者两个总体的分布不服从正态时，我们无法应用 T 检验来比较两个总体。可以转而应用非参数的方法来比较两个总体的中心位置的差异。独立样本是指样本来自的总体相互独立。

独立样本包括两个独立样本或者两个以上的独立样本。SPSS 提供的独立样本非参数检验的方法如下。

- 两个独立样本分布的比较
Mann-Whitney U
Kolmogorov-Smimov
Wald-Wolfowitz
- K 个独立样本分布的比较
Kruskal-Wallis
Jonckheere-Terpstra
- 比较全矩
Moses extreme reaction
- 比较各组的中位数
Median test

SPSS 18 的独立样本非参数检验把两个独立样本的比较和两个以上独立样本的比较放到了统一的用户界面下。

6.3.1 独立样本检验简介

SPSS 独立样本的非参数检验对话框也有和单样本的非参数检验一样的三个选项卡。

- 在“目标”选项卡上指定目标。
- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。

1. “目标”选项卡如图 6-39 所示。有如下三个目标可供选择。

- 自动比较不同组间的分布。该目标将对具有两个组的数据应用 Mann-Whitney U 检验，或对具有 k 个组的数据应用 Kruskal-Wallis 单因素 ANOVA 检验。
- 比较不同组间的中位数。该目标使用中位数检验来比较在不同组间观察到的中位数。
- 自定义分析。当您希望手动修改“设置”选项卡上的检验设置时，选中此选项。注意，如果您随后在“设置”选项卡上更改了与当前选定目标不一致的选项，则会自动选择该设置。

2. “字段”选项卡和单样本非参数检验一样。请参见 6.2 节相应部分。
3. “设置”选项卡如图 6-40 所示。它按照要进行的分析任务来组织非参数分析的方法。



图 6-39 目标选项卡

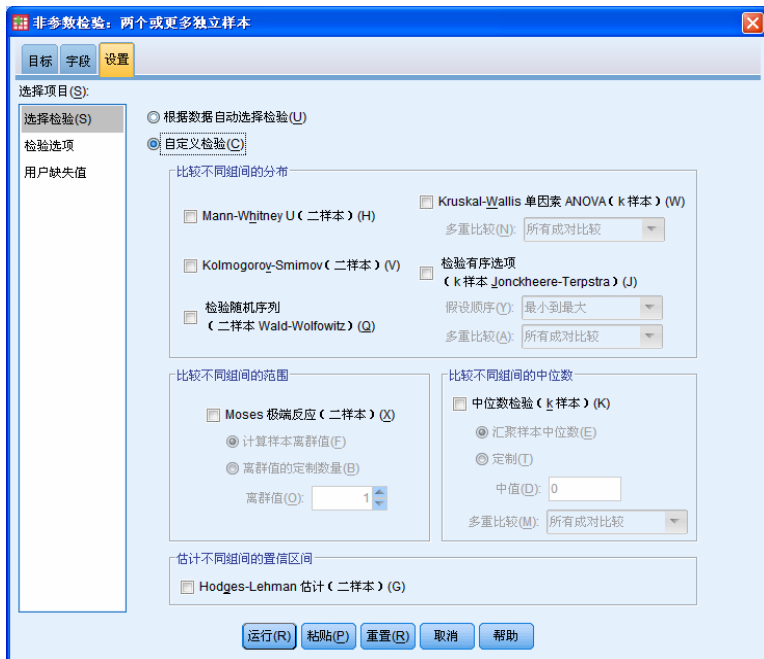


图 6-40 设置选项卡

注意：用户可根据分析目的在“自定义检验”中选择合适的非参数检验方法，当然也可“根据数据自动选择检验”，让 SPSS 自动选择所有适用的检验方法。

6.3.2 独立样本检验举例

一个公司把他们的销售代表随机分到三个不同的组中，进行不同的培训。两个月后对销售进行考察，本章的数据文件 salesperformance.sav 记录了他们的考试得分。我们想通过非参数检验比较不同组别的销售代表考试得分是否有显著性差异。这里，不同组别的考试得分是相互独立的，因此为独立样本数据，我们采用独立样本非参数检验。

打开本章的数据文件 salesperformance.sav，首先在变量视图中定义相应变量的角色。把变量“组”定义为“输入”角色，把变量“得分”定义为“目标”角色，如图 4-41 所示。

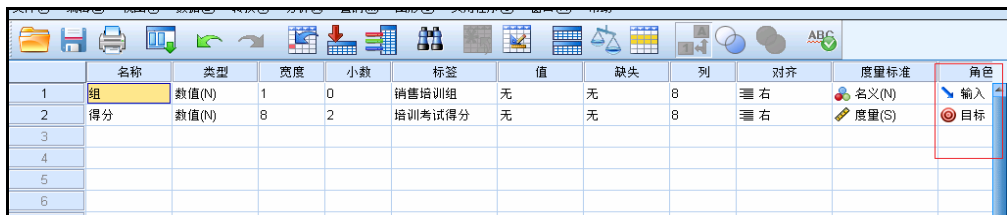


图 6-41 定义角色

然后选择【分析】→【非参数检验】→【独立样本】，如图 6-42 所示。

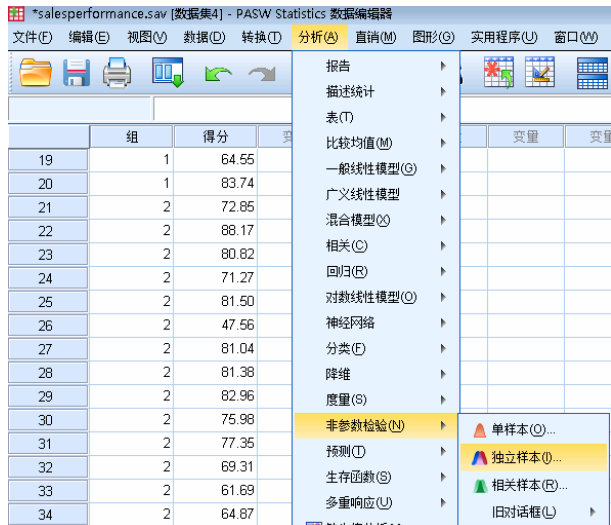


图 6-42 独立样本非参数检验

如图 6-43 所示，“字段”选项卡会自动根据之前所定义的角色分配“检验字段”

和“组”。保持默认设置，单击“运行”按钮，在输出窗口中查看分析结果，如图 6-44 到图 6-46 所示。

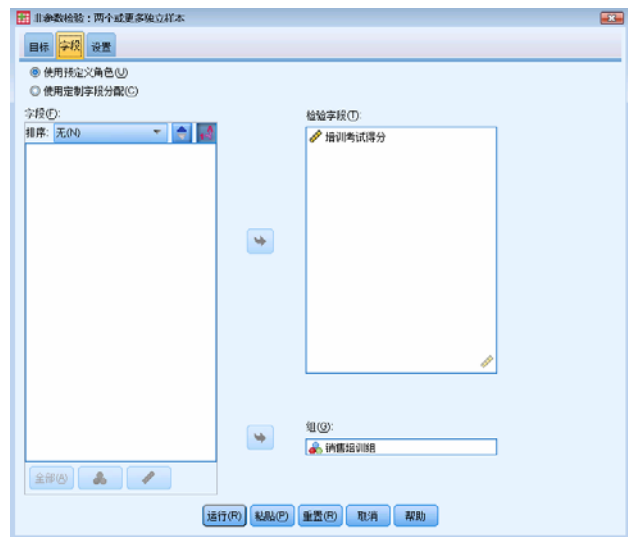


图 6-43 字段选项卡

假设检验摘要				
	原假设	检验	显著性水平	决策
1	培训考试得分 的分布在 销售培训组 的类别间相同。	独立样本 Kruskal- Wallis 检验	.000	拒绝原 假设。

显示渐近显著性。显著性水平为 .05。

图 6-44 三个独立样本非参数假设检验摘要

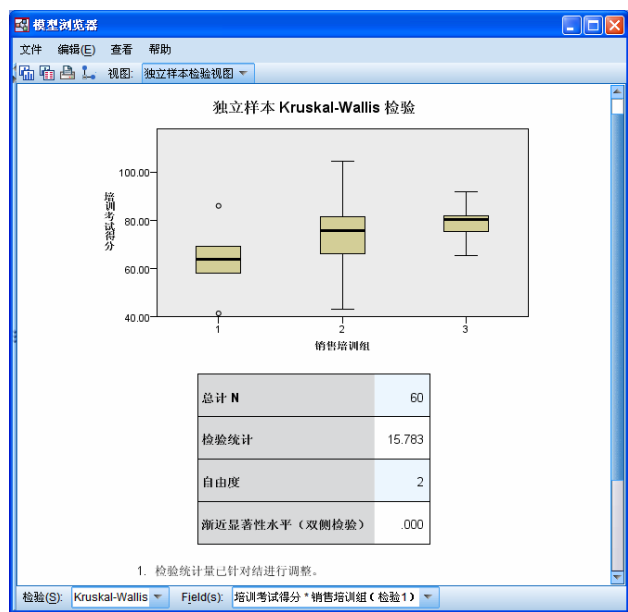


图 6-45 模型浏览器

由图 6-44 可知不同组别的销售代表得分具有显著差异，双击该图激活模型浏览器，查看更多信息，如图 6-45 所示。

图 6-45 为不同组别的箱图，从该图可知第 3 组得分最高，第 1 组得分最低。右下方为检验统计量及其 P 值。

从视图下拉框中选择“成对比较”进一步进行两组之间的“成对比较”。如图 6-46 所示。

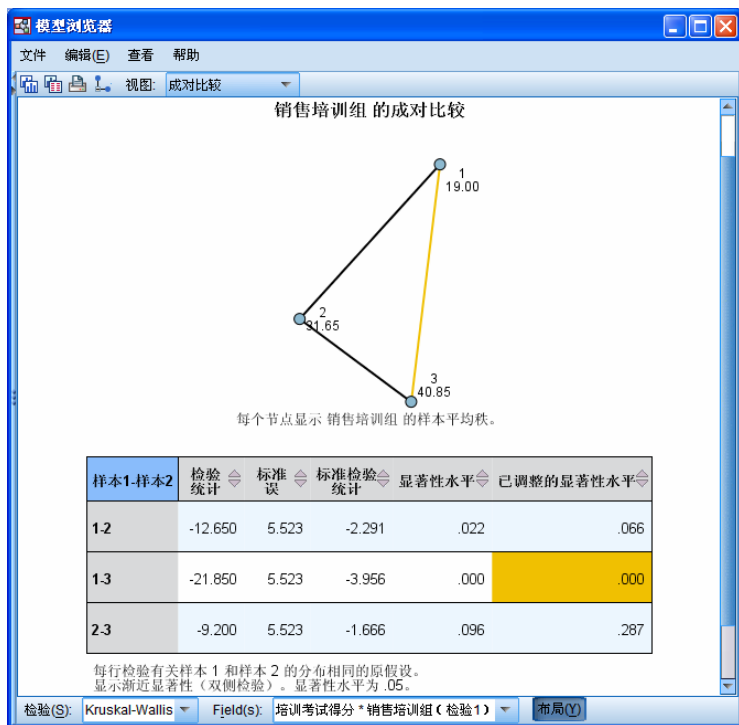


图 6-46 成对比较

从图 6-46 可直观看出：组 1 和组 3 具有显著性统计差异，而组 2 和组 3，组 1 和组 2 并无显著性统计差异。在图 6-46 上方的图中，有显著性差异的组之间连线为黄线，无显著性差异组之间的连线为黑线，图 6-46 下面的表为检验统计量和 P 值。

6.4 相关样本非参数检验

当比较一个总体的两个不同测量的差别时，如果这两个测量的分布未知，或者它们所来自的总体明显不服从正态分布时，配对的 T 检验不再适用。我们需要应用非参数的方法。SPSS 相关样本非参数检验使用一个或多个非参数检验识别两个或更多相关字段间的差别。

每个给定受试人对应一条记录，该记录包含有两个或更多相关测量值，他们分别存储在数据集的单独字段中。例如，如果每个受试人的体重以定期间隔测量并存储在如节食前体重、中间体重和节食后体重这样的字段中，则可使用样本相关非参数检验分析节食计划的有效性研究。这些字段为“相关”。

相关样本的非参数检验是配对 T 检验的推广。

6.4.1 相关样本检验简介

SPSS 相关样本的非参数检验对话框和单样本的非参数检验一样有如下三个选项卡：

- 在“目标”选项卡上指定目标；
- 在“字段”选项卡上指定字段分配；
- 在“设置”选项卡上指定专家设置。

1. “目标”选项卡如图 6-47 所示。有两个目标可供选择。

您的目标是什么？目标允许您快速指定常用的不同检验设置。

- 自动比较观察数据和假设数据。当指定 2 个字段时，该目标对分类数据应用 McNemar 检验；当指定超过 2 个字段时，则对分类数据应用 Cochran 的 Q 检验；当指定 2 个字段时，对连续数据应用 Wilcoxon 匹配对符号秩检验；当指定超过 2 个字段时，对连续数据应用 Friedman 的按秩二因素 ANOVA 检验。
- 自定义分析。当您希望手动修改“设置”选项卡上的检验设置时，选中此选项。注意，如果您随后在“设置”选项卡上更改了与当前选定目标不一致的选项，则会自动选择该设置。

当指定了不同测量级别的字段时，它们首先由测量级别隔开，然后将相应检验应用到每个组。例如，如果您选择自动比较观测数据和假设数据作为您的目标，并指定 3 个连续字段和 2 个名义字段，则会将 Friedman 检验应用到连续字段并将 McNemar 检验应用到名义字段。



图 6-47 目标选项卡

2. “字段”选项卡和单样本非参数检验一样。请参见 6.2 节相应部分。
3. “设置”选项卡如图 6-48 所示。它按照要进行的分析任务来组织非参数分析的方法。

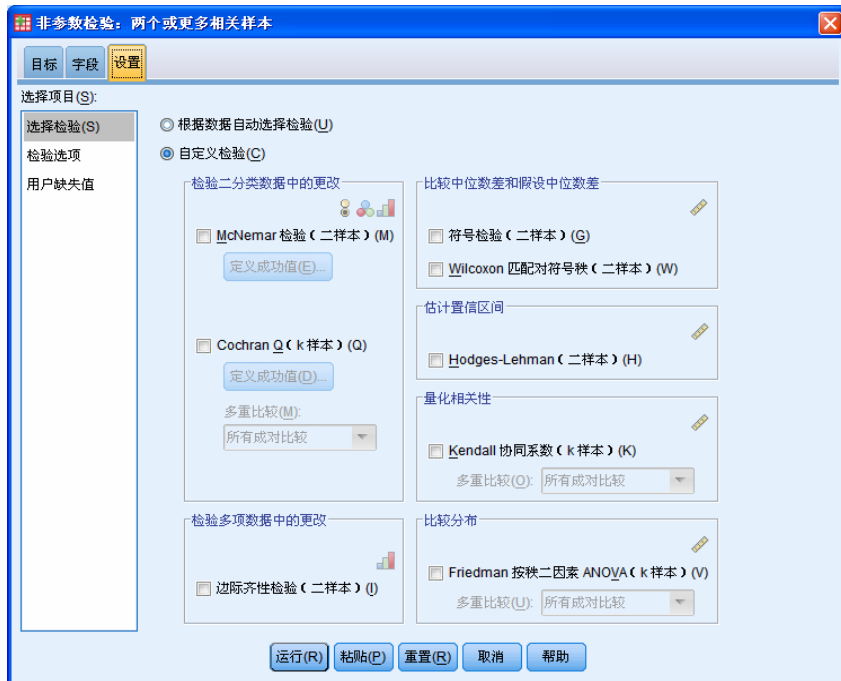


图 6-48 自定义检验方法

“自定义检验”中的设置允许您选择要执行的特定检验。

- 检验二分类数据中的更改。McNemar 检验（二样本）可以应用到分类字段。这将生成一个相关样本检验，即两个标记字段（只有两个值的分类字段）间的值组合可能性是否相同。如果在“字段”选项卡上指定两个以上的字段，将不执行此检验。Cochran 的 Q（ k 样本）可以应用到分类字段。这将生成一个相关样本检验，即 k 个标记字段（只有两个值的分类字段）间的值组合可能性是否相同。您可以根据需要请求对 k 样本的多重比较，即所有成对多重比较或逐步降低比较。
- 检验多项数据中的更改。边际齐性检验（二样本）生成一个相关样本检验，即两个配对有序字段间的值组合可能性是否相同。边际齐性检验通常在重复度量情况下使用。此检验是 McNemar 检验从二值响应到多项式响应的扩展。如果在“字段”选项卡上指定两个以上的字段，将不执行此检验。
- 比较中位数差和假设中位数差。这些检验分别生成一个相关样本检验，即两个连续字段间的中位数差是否等于 0。如果在“字段”选项卡上指定两个以上的字段，将不执行这些检验。
- 估计置信区间。这将为两个配对连续字段间中位数差生成一个相关样本估计和置信区间。如果在“字段”选项卡上指定两个以上的字段，将不执行此检验。
- 量化关联。Kendall 协同系数（ K 样本）将生成对裁判员或评分员间一致性的测量，每条记录为单个裁判员对多个项目（字段）的评价。您可以根据需要请求对 k 样本的多重比较，即所有成对多重比较或逐步降低比较。
- 比较分布。Friedman 按秩二因素 ANOVA（ k 样本）将生成一个相关样本检验，即 k 相关样本是否从同一总体中抽取。您可以根据需要请求对 k 样本的多重比较，即所有成对多重比较或逐步降低比较。

6.4.2 相关样本检验举例

本章的数据文件 `healthplans.sav` 记录了某公司雇员对四种不同医疗保险计划的评价，每个雇员对每一种医疗保险方案给出从“非常不喜欢”到“非常喜欢”四种不同评价中的一种。我们想检验公司雇员对不同医疗保险计划的喜好程度是否有显著差别。该数据为同一个雇员的四种不同评价，为相关样本数据，因此采用相关样本非参数检验。

打开本章的数据文件 `healthplans.sav`，选择【分析】→【非参数检验】→【相关样本】，如图 6-49 所示。

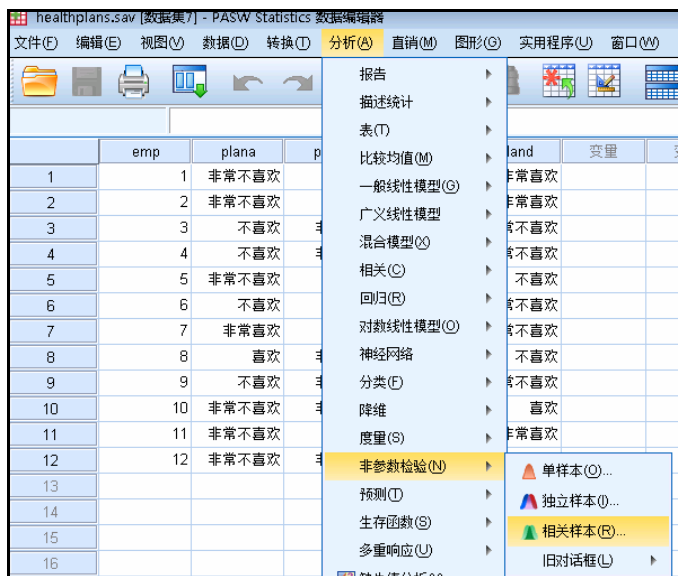


图 6-49 相关样本非参数检验

在“字段”选项卡中把 4 种医疗保险计划选入“检验字段”，如图 6-50 所示。

在“设置”选项卡中选择“Friedman 按秩二因素 ANOVA (K 样本)”，并且在多重比较中的下拉框中选择“逐步降低”，如图 6-51 所示。

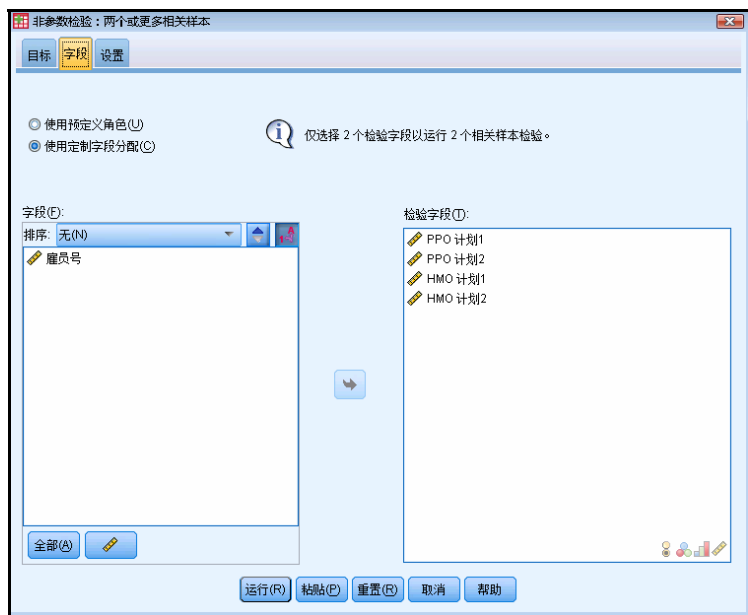


图 6-50 选择检验字段



图 6-51 检验设置

设置完毕后，单击【运行（R）】按钮。

输出的分析结果如图 6-52 到 6-54 所示。

假设检验摘要			
	原假设	检验	显著性水平
1	PPO 计划1, PPO 计划2, HMO 计划1 and HMO 计划2 的分布相同。	相关样本 Friedman 的双向按秩次方差分析	.016
			拒绝原假设。

显示渐近显著性。显著性水平为 .05。

图 6-52 假设检验摘要

从图 6-52 可知不同医疗保险计划的喜好程度具有显著性差异，双击该图激活模型浏览器以查看更详细的信息，如图 6-53 所示。

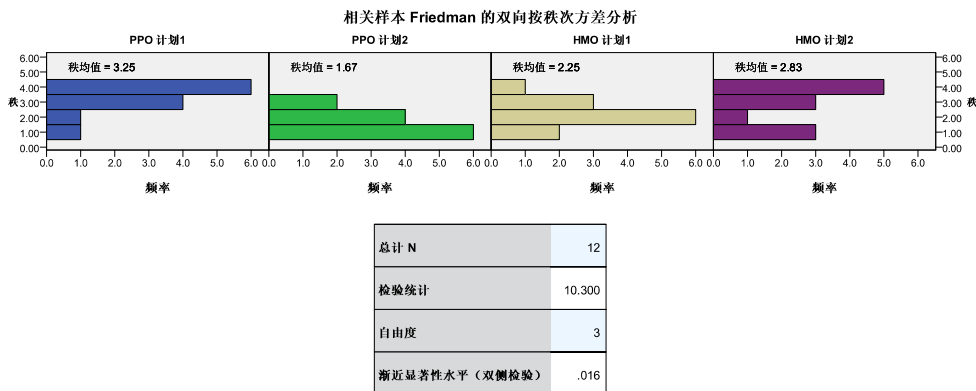


图 6-53 模型浏览器

从图 6-53 可知，PPO 计划 1 的秩均值 3.25 最高，PPO 计划 2 的秩均值 1.87 最低，图下侧为检验统计量及其相应的 P 值。

在模型浏览器的“视图”下拉框中选择“均一子集”进一步看看哪些医疗保险计划划分在同一子集中，如图 6-54 所示。



图 6-54 均一子集

从图 6-54 可直观地看出：“PPO 计划 2、HMO 计划 1、HMO 计划 2”可划分在同一子集中，同样地，“HMO 计划 1、HMO 计划 2、PPO 计划 1”也可划分在同一子集中，同一颜色用来表示同一子集。

6.5 小结

本章主要介绍了非参数检验的方法以及如何在 SPSS 中实现这些方法。当参数检验的条件不能满足时，需要采用相应的非参数的方法。非参数方法比参数方法有更广泛的应用范围，但是它的效能没有参数检验高。6.2 节为单样本的非参数检验方法，包括卡方检验、二项式检验、游程检验、K-S 检验和 Wilcoxon 检验；6.3 节介绍独立样本的非参数检验，它可以同时检验两个或者两个以上的独立样本；6.4 节介绍了相关样本的非参数检验，它可以同时检验两个或者两个以上的相关样本。

思考与练习

1. 细菌污水处理厂的微生物生态系统中最重要的一部分。水资源管理工程师认为在某个指定工厂收集的污水样本中活性细菌的百分数的中位数为 40。如果活性细菌的百分数的中位数大于 40，则应该调整污水处理过程。数据 Water.sav 记录了含有 10 个污水样品的随机样本中活性细菌的百分数。在显著性水平为 5% 的条件下，该样本提供了充分证据表明污水样本中活性细菌的百分数的中位数大于 40 吗？
2. K-S 检验可以用于检验数据是否为正态，但是在非参数检验部分的 K-S 检验和 SPSS 提供的检验数据正态性的 K-S 检验不同，请指出二者的不同指出，并比较二者的优缺点。
3. 如果读者有应用 SPSS 版本 17 或者以前版本的经验，请比较它们和 SPSS 版本 18 的非参数检验过程的异同点。你更喜欢哪一个？请说明你的理由。
4. SPSS 中进行数据的正态性检验的程序有：
 - A) Q-Q 图
 - B) P-P 图
 - C) K-S 检验
 - D) S-W 检验
5. 有关非参数检验的论断，正确的是
 - A) 非参数方法和参数方法是等价的，用户根据自己的喜好进行选择
 - B) 在可以应用参数检验的条件下，应该首先选择参数检验的方法
 - C) T 检验也可以作为非参数检验的一种方法
 - D) 以上论断都不正确

参考文献

1. 张文彤, 闫洁.SPSS 统计分析基础教程.北京: 高等教育出版社, 2004。
2. 薛薇.SPSS 统计分析方法及应用.北京: 电子工业出版社, 2009。
3. 卢淑华.社会统计学(第三版).北京: 北京大学出版社, 2005。
4. 吴喜之.非参数统计(第二版).北京: 中国统计出版社, 2006。
5. Michael Sullivan, III, Statistics, NJ: Prentice Hall,2003。

6. SPSS 18 用户手册，CD-ROM。

相关分析

本章学习目标:

- 掌握相关分析的基本概念;
- 掌握如何绘制各种散点图;
- 掌握三种相关系数的概念和解释;
- 掌握偏相关分析的概念、方法和结果解释。

7.1 相关分析的基本概念

相关分析是分析客观事物之间关系的定量分析方法。许多事物或现象之间总是相互联系的,并且可以通过一定的数量关系反映出来。比如,教育需求量与居民收入水平之间,科研投入与科研产出之间,投资额和国民收入等等,都有着一定的依存关系。而这种依存关系一般可分为两种类型:一种是函数关系,另一种是相关关系。

函数关系是指事物或现象之间存在着严格的依存关系,其主要特征是它的确定性,即对一个变量的每一个值,另一个变量都可以根据确定的函数关系取唯一确定的值与之相对应。变量之间的函数关系通常可以用函数式 $Y=f(X)$ 确切地表示出来。例如,圆的周长 C 对于半径 r 的依存关系就是函数关系: $C=2\pi r$ 。

如果我们所研究的事物或现象之间,存在着一定的数量关系,即当一个或几个相互联系的变量取一定数值时,与之相对应的另一变量的值虽然不确定,但按某种规律在一定的范围内变化。我们把变量之间的这种不稳定、不精确的变化关系称为相关关系。相关关系反映出变量之间虽然相互影响,具有依存关系,但彼此之间却没有一一对应关系。例如,学生成绩与其智力因素,人的身高和体重等。

在复杂的社会系统中,各种事物或现象之间的联系大多体现为相关关系,而不是函数关系,这主要是由于影响一个变量的因素很多,而其中一些因素还没有被人

们所完全认识和掌握,或是处于已经认识但对其产生的影响还不能完全控制和测量。另外,有些因素尽管可以控制和测量,但在操作过程中或多或少都会有误差,所有这些偶然因素的综合作用导致了变量之间的不确定性。

7.1.1 相关关系的种类

相关关系可分为线性相关和非线性相关。两个变量中的一个变量增加,另一个变量随之发生大致均等的增加或减少,它们的散点图近似地在一条直线附近,这种相关关系就称为线性相关。当两个变量中的一个变量变动时,另一个变量也相应地发生变动,但这种变动不是均等的,它们的散点图近似地表现在一条曲线附近,这种相关关系被称为曲线相关。

线性相关可分为正相关和负相关。正相关是指一个变量数值增加或减少时,另一个变量的数值也随之增加或减少,两个变量变化方向相同。负相关是指两个变量变化方向相反,即随着一个变量数值的增加,另一个变量的数值反而减少;或随着一个变量数值的减少,另一个变量数值反而增加。

根据变量的度量类型,变量相关性的研究可以分为定类变量和定类变量之间的相关,定序变量和定序变量之间的相关,尺度变量和尺度变量之间的相关。

7.1.2 相关分析的作用

在统计学中,一般将描述和分析两个或两个以上变量之间相关的性质及其相关程度的过程,称之为相关分析。相关分析的目的主要是通过具体的数量描述,呈现变量之间的相互关系的密切程度及其变化规律,找到它们相互关联的模式,从而根据此模式做出决策或者为进一步采取其他统计分析手段提供参考依据。

相关分析在统计分析中的作用是多方面的,具体概括如下。

1. 判断变量之间有无联系

确定研究现象之间是否具有依存关系,这是相关分析的起点,也是我们研究各种现象之间相互关系的前提条件。因为只有确定了依存关系的存在,才有继续研究和探索各种现象之间相互作用、制约以及变化规律的必要和价值。

2. 确定相关关系的表现形式及相关分析方法

在确定了变量之间存在依存关系之后,就需要明确体现变量相互关系的具体表

现形式。在此基础上，选择恰当的相关分析方法，只有这样才能确保研究目的的实现，收到预期的效果。否则，如果把非线性相关错判为线性相关，按照线性相关的性质选择相关分析的方法，就会导致错误的结论。

3. 把握相关关系的方向与密切程度

变量之间的相关关系是一种不精确的数量关系，相关分析就是要从这种不确定的数量关系中，判断相关关系的方向和密切程度。

4. 为进一步采取其他统计方法进行分析提供依据。

本质上，相关分析是一种探索性的统计分析方法，根据其结果和分析的目的，可以为下一步的分析提供指导。如果相关分析的结果是两个尺度变量间有较强的线性关系，那么线性回归可能是下一步的分析方法；如果是非线性关系，那么下一步可以考虑非线性回归、曲线拟合等方法。

5. 相关分析不但可以描述变量之间的关系状况，而且可以用来进行预测。

另外，相关分析还可以用来评价测量量具的信度、效度以及项目的区分度等。

7.2 散点图

进行相关分析的主要方法有图示法和计算相关系数法。图示法主要是通过绘制相关散点图，找出变量之间相关关系的模式的方法。它是一种探索性分析的方法，需要和相应的相关系数结合来进行分析和判断。计算相关系数法则是根据不同类型的数据，选择不同的计算方法求出相关系数，据此来进行相关分析的方法，我们将在第 7.3 节和 7.4 节介绍。

7.2.1 散点图简介

相关散点图是观察两个变量之间关系的一种非常直观的方法。散点图以横轴表示两个变量中的一个变量，以纵轴表示另一个变量，将两个变量之间相对应的变量值以坐标点的形式逐一标在直角坐标系中，通过点的分布形状、分布模式和疏密程度来形象描述两个变量之间的相关关系。SPSS 的图表构建程序和旧统计图对话框都可以完成散点图的绘制。

本章的数据文件 `car_sales.sav` 记录了对市面上常见汽车的调查结果，它包括车的长、宽、净重等物理指标，同时还有车的厂家、型号、新车售价、发动机、马力、

耗油量等。我们想考察车的耗油量是否和售价有关系，是否车越省油价格越高呢？

首先画出新车售价和耗油量的散点图。

7.2.2 散点图——旧对话框

打开数据文件 car_sales.sav，选择【图形】→【旧对话框】→【散点/点状(S)】，如图 7-1 所示，出现散点图选择对话框，如图 7-2 所示。

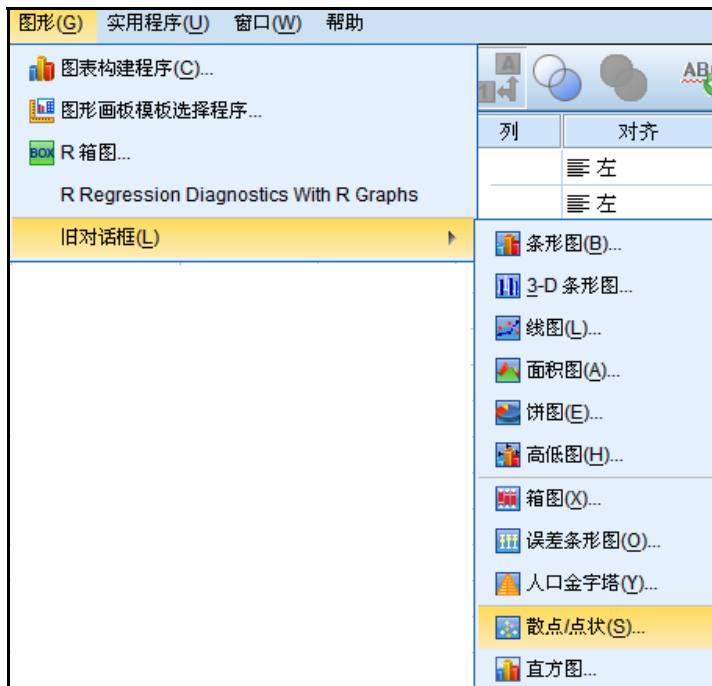


图 7-1 散点图——旧对话框



图 7-2 选择散点图类型

这里选择第一个——“简单散点图”，得到“简单散点图”对话框，如图 7-3 所示。该对话框定制散点图的两个变量，点子的标识方式等。

- “X轴”和“Y轴”分别设置散点图的横轴和纵轴所代表的变量，原则上哪个变量作散点图的X轴或者Y轴没有硬性规定，两个变量都既可以作X轴，也可以作Y轴。这里，要考察的两个变量分别为耗油量和新车售价，我们把mpg选入X轴框中，把sales选入Y轴框中。
- “设置标记”将在散点图中通过图例的方式来标识散点图中的点。我们这里保留默认值。
- “标注个案”将给散点图上的点添加文字标识。我们代表汽车型号的变量model选入该框中。
- 面板依据：用于设置多组散点图。这里我们保留默认值。

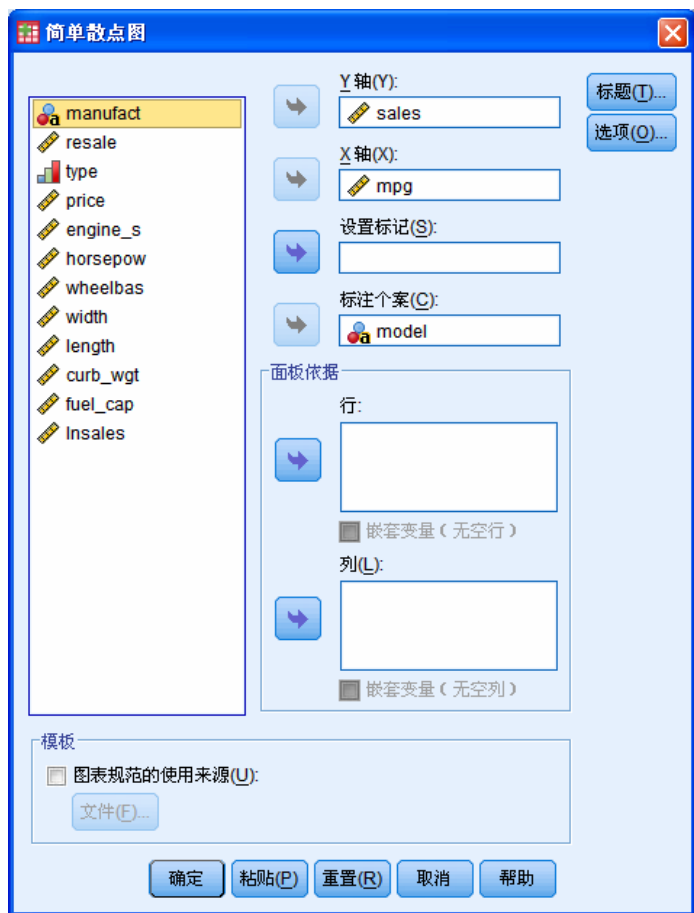


图 7-3 简单散点图

单击【确定】按钮。

以上操作可以通过以下语法程序来完成。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 7\car_sales.sav' .
```

```

DATASET NAME myData WINDOW=FRONT.
GRAPH
  /SCATTERPLOT(BIVAR)=mpg WITH sales BY model (IDENTIFY)
  /MISSING=LISTWISE.
EXECUTE

```

结果浏览器中得到散点图如图 7-4 所示。

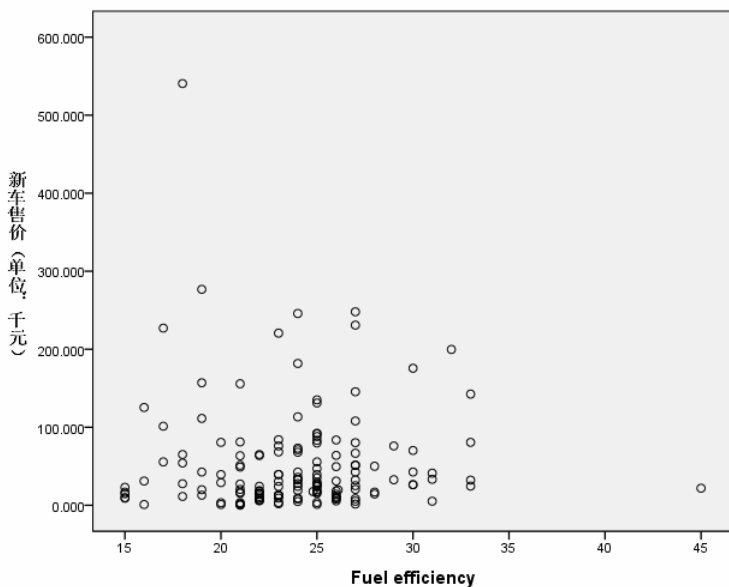


图 7-4 简单散点图

从图 7-4 所示的简单散点图可知，新车售价和耗油量之间的关系不明显，事实是这样吗？我们在后面的几节中将进一步分析这两个变量之间的关系。注意到图 7-4 的散点图中左上和右下的两个点偏离了大部分的点，我们称它们为离群值（Outliers）。那么，这两个点对应哪种型号的车呢？在分析中是否应该把它们去掉呢？

可以通过显示散点图中点子的标签来找出它们对应的汽车的型号。在结果查看器中双击图 7-4，进入图表编辑器。然后选择【元素】→【显示数据标签】，散点图中的点将被附以它们对应的车型号，如图 7-5 所示。

关闭如图 7-5 所示的图表编辑器后，结果查看器中的散点图将有点子的标签显示，如图 7-6 所示。

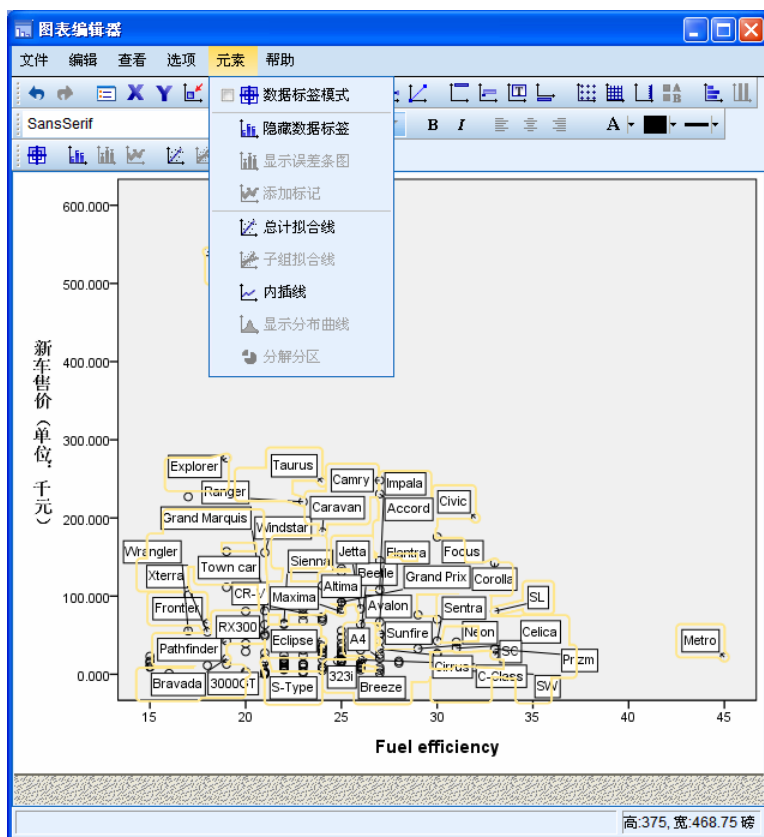


图 7-5 图表编辑器

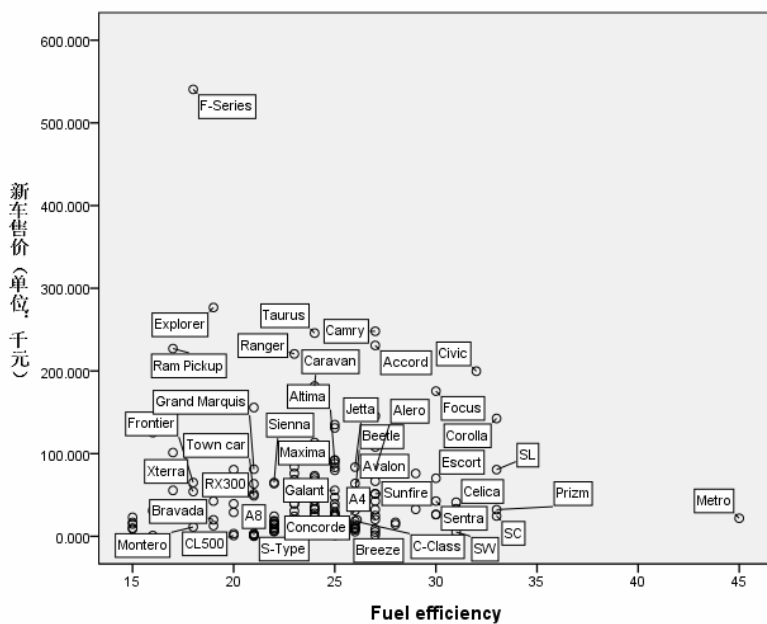


图 7-6 附加标签的散点图

从图 7-6 可以看出，左上交的车型为 F-Series，右下角的车型为 Metro。如果

F-Series 是某公司正在开发的主力车型，而 Metro 是比较老的一款车型，目前不是销售主流，那么在后续的分析中可以保留 F-Series 相对应的个案而去掉 Metro 所对应的个案。保留了左上的 F-Series 点，新车售价的分布将呈偏态分布。我们对新车售价取自然对数后来代替原来的新车售价变量进行后续分析。

7.2.3 用图表构建程序绘制散点图

SPSS 集成所有的图形程序于图表构建程序中，所有的统计图形的绘制有统一的用户界面。尽管 SPSS18 仍然保留了旧对话框以和以前的版本兼容，图表构建程序将是 SPSS 统计图重点发展的部分。

选择【图形】→【图表构建程序(C)】，如图 7-7 和图 7-8 所示。

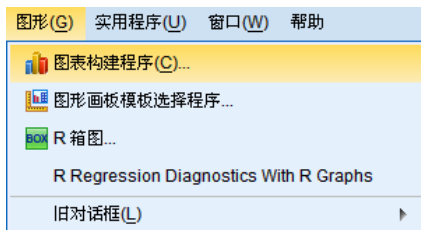


图 7-7 图表构建程序

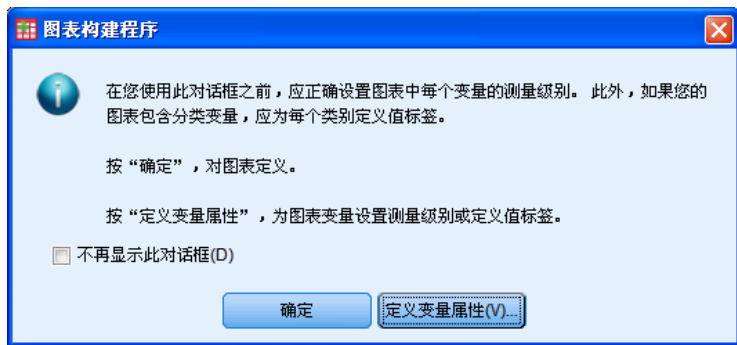


图 7-8 提示

图 7-8 所示的对话框用于提示正确设置变量的测量级别。可以勾选“不再显示此对话框”，以后启动图表构建程序时将不在显示该提醒。单击【确定】按钮，如图 7-9 所示。



图 7-9 图表构建程序

首先在“库”标签部分，选择“散点图/点图”，然后用鼠标把左上的第一个图拖放到对话框右上的空白中。然后选择“组/点 ID”标签，勾选“指定 ID 标签”。最后把相应的变量按照图 7-9 所示选入到相应的位置，单击【确定】按钮。

结果查看器中得到和图 7-6 完全类似的带标签的散点图，如图 7-10 所示。

以上操作过程可以通过以下语法命令完成。

```

DATASET ACTIVATE myData.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=mpg sales model
  MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: mpg=col(source(s), name("mpg"))
  DATA: sales=col(source(s), name("sales"))
  DATA: model=col(source(s), name("model"), unit.category())
  GUIDE: axis(dim(1), label("耗油量:迈/升"))
  GUIDE: axis(dim(2), label("新车售价(单位:千元)"))

```

```
ELEMENT: point(position(mpg*sales), label(model))
END GPL.
```

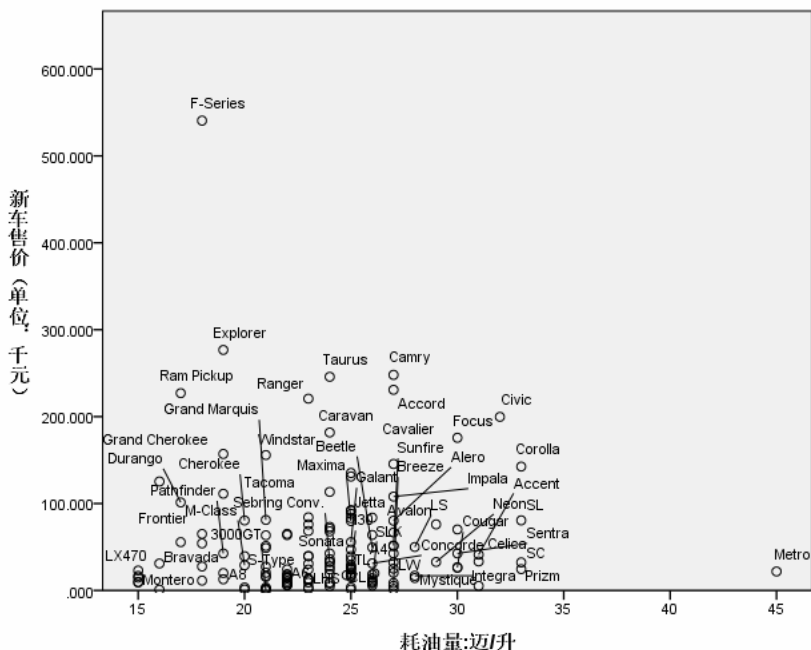


图 7-10 带标签的散点图

7.3 相关系数

计算法就是通过计算相关系数来分析变量之间相互关系的方法。计算相关系数的方法很多，由于我们所面对的各种变量都具有不同的性质和类型，因此应当根据变量的特点选择适当的分析相关的方法。对于不同类型的数据，计算相关系数的方法也不相同。

下面介绍几种适用于不同类型的变量相关分析的计算方法。

7.3.1 线性相关的度量——尺度数据间的相关性的度量

假设我们有两个变量 X 和 Y ，我们用协方差和线性相关系数来衡量这两个变量线性相关的程度。

1. 协方差

假设我们有两个连续变量 X 和 Y ，它们的观测值分别为 $x_i, y_i, i = 1, \dots, n$

它们的样本均值为 $\bar{y} = \sum_{i=1}^n y_i / n$ 和 $\bar{x} = \sum_{i=1}^n x_i / n$

样本标准差分别为 $s_x = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)}$ 和 $s_y = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)}$

那么协方差为:

$$\text{Cov}(X, Y) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n-1)$$

协方差为 0 表明两个变量之间没有线性关系。协方差为正说明两个变量之间有正线性相关关系, 为负值说明两个变量之间有负线性相关关系。协方差和变量的量纲有关, 其符号表明线性关系的方向, 但是协方差的大小它不能表明两个变量之间关系的强弱。

2. Pearson 相关系数

相关系数克服了协方差和量纲有关的缺点, 它即可以衡量两个变量是否有线性相关关系, 同时在线性相关条件下, 可以描述两个变量之间线性相关的方向和相关的程度。相关系数由 Pearson 最早提出, 又称为 Pearson 相关系数。其定义如下。

$$\rho = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{x_i - \bar{x}}{s_x} \right) = \frac{\text{Cov}(Y, X)}{s_y s_x}$$

相关系数的数值范围是介于 -1 与 +1 之间:

如果 $|\rho| \approx 0$, 表明两个变量没有线性相关关系。

如果 $|\rho| \approx 1$, 则表示两个变量完全直线相关。线性相关方向通过相关系数的符号来表示, “+” 号表示正相关, “-” 表示负相关。

注意:

- 相关系数为 0 或接近于 0 不能说明两个变量之间没有相关性, 它只说明没有线性相关性。不能排除具有其他非线性关系。
- Pearson 相关系数是一种线性关联度量。如果两个变量关系密切, 但其关系不是线性的, 则 Pearson 相关系数就不是适合度量其相关性的统计量。
- Pearson 相关系数适用于两变量的度量水平都是尺度数据, 样本量大于 30, 并且两个变量的总体是正态分布或近似正态分布的情况, 否则其反映的线性关系有失真的可能。

3. 计算 Pearson 相关系数

SPSS 的双变量相关可以计算两个或者两个以上变量间的协方差和 Pearson 相关系数。同时还可以检验该相关系数是否显著区别于 0。设相关系数为 ρ ，则 SPSS 相关系数检验的原假设为：

$$H_0: \rho = 0$$

如果计算得到的相关系数显著，即 H_0 不成立，SPSS 将在该相关系数的右上角标注(**)。

本章的数据文件 car_sales.sav 记录了对市面上常见汽车的调查结果，它包括车的长、宽、净重等物理指标，同时还有车的厂家、型号、新车售价、发动机、马力、耗油量等。我们想考察车的耗油量是否和新车售价有关系，是否车越省油价格越高呢？

在 7.2 中，我们绘制了车的耗油量和新车售价的散点图，下面我们计算二者之间的相关系数。选择【分析】→【相关】→【双变量】，得到图 7-11。把 sales 和 mpg 选入“变量 (V)”框中，如图 7-11 所示。

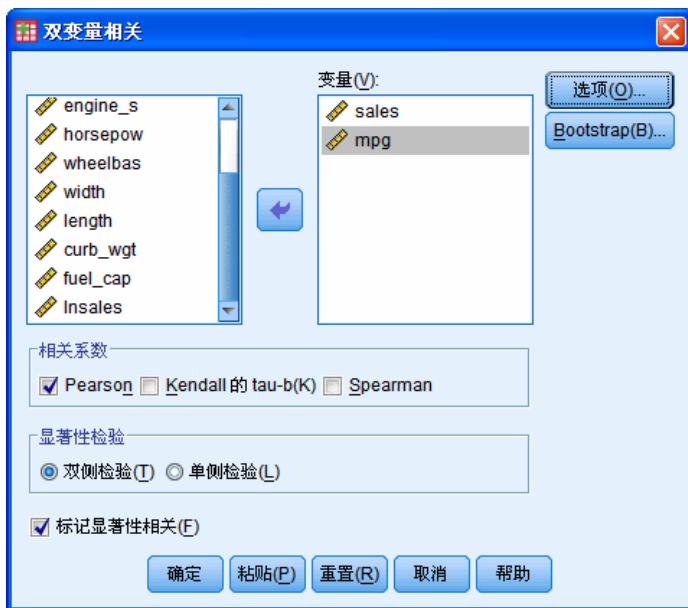


图 7-11 双变量相关

图 7-11 中的选项简介如下：

- “相关系数”：指定相关系数的类型；

- “显著性检验”：指定对相关系数检验的类型，可以选择双尾概率或单尾概率。如果预先已知关联的方向，选择单侧检验。否则，选择双侧检验。
- “标记显著性相关(F)”：用(*)来标识在显著性水平 0.05 下显著的相关系数；用(**)来标识显著性水平 0.01 下显著的相关系数。

如图 7-11 设置完毕后，单击【确定】按钮。

以上操作可以通过下列的语法命令来完成。

```
CORRELATIONS
/VARIABLES=sales mpg
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

在结果查看器中得到表 7-1 所示的结果。从中可以看出新车售价和耗油量的相关系数为-0.017，该相关系数的显著性值为 0.837，大于 0.10。即新车售价和耗油量之间的线性相关不显著。这说明汽车的耗油量不是影响汽车销售的一个重要因素，汽车的设计者不必在使汽车更加省油上花费更大的精力。Pearson 相关系数在变量分布呈正态或者近似正态并在没有离群值的情况下能更好地揭示变量之间的线性关系。从 7.2.2 节的图 7-6 可知，我们分析的样本数据有离群值。另外，是否对不同类型的车—轿车和卡车，售价和耗油量之间的线性相关关系都不显著呢？

表 7-1 相关性

		新车售价(单位：千元)	耗油量:迈升
新车售价(单位：千元)	Pearson 相关性	1	-.017
	显著性(双侧)		.837
	N	157	154
耗油量:迈升	Pearson 相关性	-.017	1
	显著性(双侧)	.837	
	N	154	154

我们先从数据集中剔除离群值 Metro 所对应的记录，然后分析售价和耗油量在不同类型的车中的线性相关系数。

1) 剔除离群值

选择【数据】→【选择个案】，如图 7-12 所示。

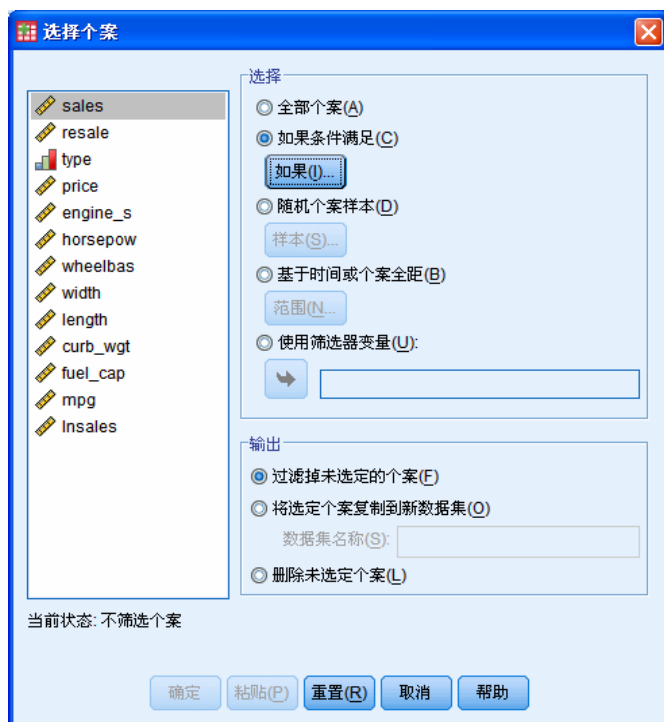


图 7-12 选择个案——剔除离群值

在图 7-12 中，单击，“如果(I)”，得到如图 7-13 所示的对话框。在右上的文本框中，我们输入“model<=>'Metro'”。



图 7-13 输入选择条件

单击【继续】，返回上级对话框，然后单击【确定】按钮。以后的分析中'Metro'所在的个案将不在被分析的数据中。

2) 文件分割

我们按照车的类型—type 进行文件分割。选择【数据】【拆分文件】，得到如图 7-14 所示的分割文件对话框。



图 7-14 分割文件

3) 计算相关系数

我们重新计算新车售价和耗油量两个变量的 Pearson 线性相关系数。得到结果如表 7-2 所示。

表 7-2 按照车类型的相关系数

Vehicle type			耗油量/迈升	新车售价(单位：千元)
轿车	耗油量/迈升	Pearson 相关性	1	.328**
		显著性(双侧)		.000
		N	113	113
	新车售价(单位：千元)	Pearson 相关性	.328**	1
		显著性(双侧)	.000	
		N	113	115
卡车	耗油量/迈升	Pearson 相关性	1	.011
		显著性(双侧)		.944
		N	40	40
	新车售价(单位：千元)	Pearson 相关性	.011	1
		显著性(双侧)	.944	
		N	40	41

**在 .01 水平(双侧)上显著相关。

从表 7-2 可知，对于卡车，耗油量和新车售价相关系数为 0.011，其显著性值为

0.944, 大于 0.1, 即对于卡车, 设计人员的确不用花太大精力在降低耗油量上; 而对于轿车, 新车售价和耗油量的 Pearson 相关系数为 0.328, 其显著性值为 0.000, 即对于轿车, 其售价是和耗油量正线性相关的, 要提高新车的售价, 设计人员需要提高车的单位油耗的里程数。

注意:

- 由于离群值的存在, 新车售价的分布是偏态的, 可以通过取自然对数, 使得变量的分布更接近于正态, 这时取过对数的售价和耗油量两个变量间的线性相关系数更好地揭示二者之间的线性关系。统计上有较多的方法来分析线性模型, 并且这些模型相对易于实现和理解。读者可以自己分析取过自然对数的售价($\ln \text{sale}$)和耗油量间的关系。
- Spearman 相关系数和 Kendall 的 τ -b 相关系数和数据的分布无关, 它们只考察两个变量的秩次相关性, 同时它们对离群值不敏感。对于偏态较大的尺度型数据, 可以考虑应用这两个系数来考察相关性。Spearman 相关系数应用范围比 Pearson 相关系数广, 但是其统计效能比 Pearson 相关系数要低一些 (不容易检测出两者事实上存在的相关关系)。读者请自己试验。
- SPSS 的双变量相关同时可以给出协方差。这通过设置图 7-11 的选项即可。

7.3.2 Spearman 等级相关系数一定序变量之间的相关性的度量

在进行相关分析的过程中, 我们经常会遇到一些不适宜应用 Pearson 相关系数的具有等级顺序的测量数据, 在这种情况下, 要研究两个或两个以上变量的相关, 就需要采用等级相关。这种相关方法对变量的总体分布不做要求, 因此又称这种相关为非参数相关。SPSS 提供的计算等级相关的方法有 Spearman 等级相关和 Kendall 的 τ -b 相关。

当两列变量值为定序数据, 并且变量值所属的两个总体并不一定呈正态分布, 样本容量也不一定大于 30, 这两个变量之间的相关性可以通过计算 Spearman 等级相关系数进行分析。由于这种相关是英国统计学家 Spearman 根据 Pearson 相关公式推导得到的, 某种意义上也可以认为 Spearman 等级相关是 Pearson 相关的一种特殊形式, 它不是对变量 X 和 Y 的值应用 Pearson 相关系数公式, 而是对变量 X 和 Y 的秩应用 Pearson 相关系数公式。

斯皮尔曼等级相关的适用条件为:

- 两个变量为定序变量;

- 一个变量为定序变量，另一个变量为尺度数据，且两总体不是正态分布，样本容量 n 不一定大于 30。

从 Spearman 等级相关适用条件中可以看出，等级相关的应用范围要比 Pearson 相关广泛，它的突出优点是对数据的总体分布、样本大小都不做要求，对离群值不敏感。但缺点是效能不高。另外，如果一个变量的某个值对应另一个变量的若干个不同的取值的个案较多时，Spearman 相关系数则不宜使用。Spearman 等级相关系数常用符号 ρ 来表示，其计算公式为：

$$\rho = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)} \quad (\text{Spearman 相关系数公式})$$

其中， D 是两个变量每对数据的等级差， n 是样本量。例如，表 7-3 记录了每周看电视的时间和 IQ 之间的关系，我们用 Spearman 等级相关分析二者的相关性。

表 7-3 Spearman 相关

智商	每周看电视时间	智商等级	每周看电视时间等级	等级差 d_i	等级差平方 d_i^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

根据 Spearman 相关系数公式， $\sum_{i=1}^{10} d_i^2 = 0 + 16 + 25 + 9 + 25 + 9 + 16 + 9 + 49 + 36 = 194$

$$\rho = 1 - \frac{6 \times 194}{10(10^2 - 1)} = -0.176$$

在 SPSS 中，我们可以通过选择【分析】→【双变量相关】，进入如图 7-11 所示的双变量相关对话框，在“相关系数”部分去掉 Pearson 前面的勾，同时勾选 Spearman。我们得到 Spearman 相关系数的结果如表 7-4 所示。

表 7-4 Spearman 相关系数

			智商	看电视时间
Spearman 的 rho	智商	相关系数	1.000	-.176
		Sig. (双侧)	.	.627
		N	10	10
	看电视时间	相关系数	-.176	1.000
		Sig. (双侧)	.627	.
		N	10	10

7.3.3 Kendall 的 tau-b(K)

Kendall 的 tau 系数是另一种计算定序变量之间或者定序和尺度变量之间相关系数的方法。Spearman 的等级相关系数可以方便检验两个定序变量是否相关，但是很难具体解释两个变量如何相关及相关程度。Kendall 的等级相关系数可以同时反映两个变量的相关程度。

设样本量为 n ，考察两个变量 X 和 Y 之间的相关关系， X 和 Y 的取值记为 $x_i, y_i, i=1, \dots, n$ 。所有像 $(x_i, y_j), i, j=1, \dots, n$ 对的个数为 $n(n-1)/2$ 。 $s_i, i=1, \dots, n$ 和 $t_i, i=1, \dots, n$ 分别表示 $x_i, i=1, \dots, n$ 和 $y_i, i=1, \dots, n$ 的秩次，如果对于任意 $k, k=1, \dots, n$ ，有 $s_k \geq t_k$ 我们称 (x_k, y_k) 为同序对；否则，称为逆序对。总的同序对的个数记为 n_c ，逆序对的个数记为 n_d ，则 Kendall 的 Tau 系数的定义为：

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

如果数据中的某个变量有太多的相同值，则采用修正的 tau 系数，称为 tau (b)，其公式为：

$$\tau = \frac{n_c - n_d}{\sqrt{0.5n(n-1) - T_x} \sqrt{0.5n(n-1) - T_y}}$$

SPSS 中计算 Kendall 的操作和 Pearson 或 Spearman 类似，请参照 7.3.1 节和 7.3.2 节。

7.4 偏相关分析

政府医疗基金的投入和发病率之间存在关系吗？尽管您可能希望存在一个负相关的关系，但是它们之间的相关系数表明二者存在显著的正相关关系，即随着医疗基金的增长，发病率也表现为增长。不过，对保健提供商的拜访率的控制，实际上消除了所观察到的正相关。保健基金和发病率显示为正相关的原因仅仅是，当基金

增长时，更多的人可以获得保健服务，从而导致医生和医院所报告的病例更多。

SPSS 的“偏相关”过程计算偏相关系数，该系数在控制一个或多个附加变量效应的同时描述两个变量之间的线性关系。

打开本章的数据文件 health_funding.sav 数据文件，选择【分析】→【相关】→【偏相关】，得到如图 7-15 所示的偏相关对话框。



图 7-15 偏相关对话框

把变量 funding 和 disease 选入到“变量（V）”框中，然后把 visits 选入到“控制（C）”框中。“显著性检验（T）”框中可以选择双尾概率或单尾概率。如果预先已知关联的方向，选择单尾。否则，选择双尾。

“显示实际显著性水平（D）”：默认情况下，该选项被选中，它将显示每个相关系数的概率和自由度。如果取消选择此项，则使用（*）标识显著性水平为 0.05 的系数，使用（**）标识显著性水平为 0.01 的系数，而不显示自由度。该设置同时影响偏相关矩阵和零阶相关矩阵。

【选项】按钮选择需要显示的统计量和缺失值的处理方式。零阶相关系数为所有变量（包括控制变量）之间的简单相关矩阵。这里保留默认值，如图 7-16 所示。

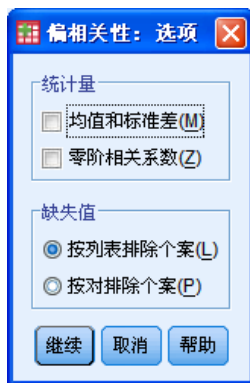


图 7-16 偏相关选项

单击【继续】按钮，返回上级菜单，单击【确定】按钮。以上操作可以通过下列语法命令完成。

```
PARTIAL CORR
/VARIABLES=funding disease BY visits
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE.
```

表 7-5 偏相关系数

控制变量			医疗基金	每万人报告病例
每万人访问医院人数	医疗基金	相关性	1.000	.013
		显著性（双侧）	.	.928
		df	0	47
每万人报告病例	每万人报告病例	相关性	.013	1.000
		显著性（双侧）	.928	.
		df	47	0

从表 7-5 可知，实际上医疗基金和报告的病例人数之间的关系不显著。

7.5 小结

本章学习相关性的概念及在数据分析中的应用。7.1 节介绍了相关分析的种类和作用，7.2 节和 7.3 节介绍相关分析的方法：散点图法和相关系数法。7.3 节学习了三种常见的描述变量间相关性强弱的指标：Pearson 线性相关系数、Spearman 相关系数和 Kendall 相关系数。最后，介绍了偏相关分析的方法，它是一种分析剔除其他变量影响后的两个变量之间关系的一种方法。

思考与练习

1. 数据文件 tourist.sav 记录了 2000 年至 2008 年入境游旅客的人数和相应年份

的收入，二者之间有关系吗？画出散点图，并求出相关系数。然后给出合理的解释。

2. 分析数据 `car_sales.sav` 中变量汽车销量和汽车耗油量之间的关系。它们是否有线性相关性？如果没有线性相关性，二者之间有其他关系吗？
3. 糖尿病病人需要靠胰岛素来治疗。数据文件 `Parcorr.sav` 记录了 20 名糖尿病病人血液中的血糖值、胰岛素值和生长激素值的结果，三者之间是否有相关性？用适当的相关分析程序来找出三者之间的关系。
4. 对于两个尺度型变量，考察他们的相关系数时，应该计算：
 - A) Pearson 相关系数
 - B) Kendall's tau-b 系数
 - C) Spearman 相关系数
 - D) 偏相关系数
5. 在 SPSS 的【分析】→【描述统计】→【交叉表】中，可以进行相关性分析，以下论断错误的是：
 - A) 可以分析名义变量之间的相关性
 - B) 可以分析定序变量之间的相关性
 - C) 可以提供卡方值
 - D) 以上论断都不正确
6. 哪些功能是 SPSS 的相关分析过程不能提供的：
 - A) 计算定量变量间的 Pearson 相关系数
 - B) 判断定量变量相关系数是否显著区别于 0
 - C) 判断相关性的强弱
 - D) 提供变量是否有非线性的相关关系

参考文献

1. 卢淑华. 社会统计学（第三版）. 北京：北京大学出版社，2005。
2. 吴喜之. 统计学：从概念到数据. 北京：高等教育出版社，2008。
3. Michael Sullivan, III, Statistics, NJ: Prentice Hall, 2003。

本章学习目标:

- 掌握简单线性回归分析的基本概念;
- 掌握线性回归的前提条件并能进行验证;
- 掌握线性回归分析结果的解释;
- 能够用线性回归模型进行预测。

8.1 线性回归分析的基本概念

回归分析是研究变量之间相关关系的一种统计方法。在第 7 章相关分析中, 如果两个变量之间的 Pearson 相关系数绝对值较大, 且变量间线性关系显著, 那么下一步就是应用回归分析的方法来找出变量之间的线性关系。

例如, 房屋的价格和房屋的面积, 地理位置, 房龄和房间的个数都有关系。又比如, 香烟的销量和许多地理和社会经济因素有关, 像消费者的年龄, 教育, 收入, 香烟的价格等。一般上述关系是以下列方程来表示

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon \quad (8-1)$$

上式中 Y 被称作因变量, 或者响应变量; 而 X_1, X_2, \dots, X_p 称作自变量、控制变量、解释变量或者预测变量; 而 f 则称为回归函数, ε 为随机误差或随机干扰, 它是一个分布与自变量无关的随机变量, 我们常假定它是均值为 0 的正态变量。

根据回归函数 f 的形式, 回归分析可以分为线性回归和非线性回归;

- 线性回归的形式为:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (8-2)$$

这里, $\beta_0, \beta_1, \dots, \beta_p$ 被称作回归系数, 它们为待定常数, ε 为随机误差。 β_0 为常数项, 有时候称为截距。对于有一个响应变量的线性回归, 当 $p=1$ 时, 我们称为

简单线性回归(Simple Linear Regression, 或称为一元线性回归), 当 $p \geq 2$ 时我们称为多元线性回归(Multiple Linear Regression)。SPSS 的“回归”菜单的“线性”子菜单可以进行简单线性回归和多元线性回归。

线性回归一般放到一般线性模型的框架下来讨论。这里的线性指的是回归系数为线性, 而非相应变量和预测变量的关系。例如方程 $Y = \beta_0 + \beta_1 \log X + \varepsilon$ 仍然认为是线性回归方程

- 非线性回归

如果预测变量和响应变量之间有公式 8.1 所示的关系, 但是不能表示为 8.2 所示的线性方程的形式, 我们称该回归关系为非线性回归。SPSS 的回归菜单下有“非线性”、“二元 Logistic”、“多元 Logistic”“有序”回归和“Probit 回归”“部分最小平方”等非线性回归程序, 另外如果安装了 SPSS 的 R 插件, SPSS 回归菜单中将可以实现“Tobit 回归”、“稳健回归”、“分位数回归”等, 如图 8-1 所示。

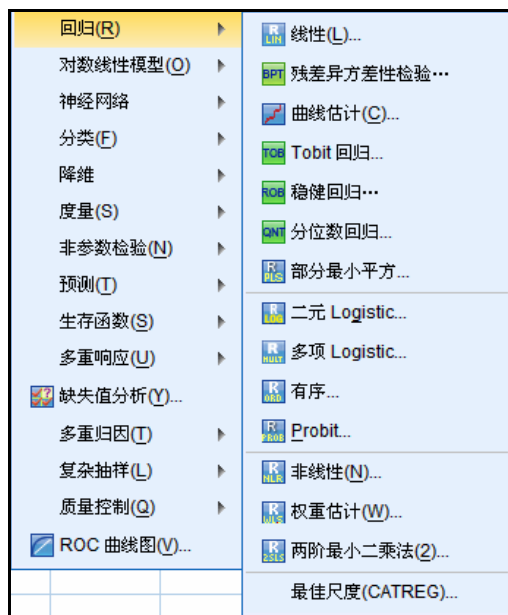


图 8-1 SPSS 回归分析菜单

回归分析是在相关分析的基础上, 确定了变量之间的相互影响关系之后, 准确的描述这种关系的数量方法。因此, 一般情况下, 相关分析要先于回归分析进行, 确定出变量间的关系是线性还是非线性, 然后应用相关的回归分析方法。在应用回归分析之前, 散点图分析是常用的探索变量之间相关性的方法。

注意：如第7章所述，可以采用散点图和相关系数进行相关分析，一般二者结合进行。由于 Pearson 相关系数受数据分布和离群值的影响，仅仅采用 Pearson 相关系数可能会有误导。

应用回归分析一般遵循下列步骤：

- 步骤 1：写出研究的问题和分析目标；
- 步骤 2：选择潜在相关的变量；
- 步骤 3：收集数据；
- 步骤 4：选择合适的拟合模型；
- 步骤 5：模型求解；
- 步骤 6：模型验证和评价；
- 步骤 7：应用模型解决研究问题。

以上有些步骤可以跳过，有些需要重复进行。例如，如果数据已经有了，那么前三个步骤可以省略；而如果步骤 6 中的模型验证的结果不满意，则需要重新进行步骤 4 到步骤 6 的过程。

8.2 简单线性回归

在简单线性回归中，只有两个变量，其回归方程为：

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (8-3)$$

其中预测变量 X 为预测变量，它是可以观测和控制的； Y 为因变量或响应变量，它为随机变量； ε 为随机误差。

通常假设 $\varepsilon \sim N(0, \sigma^2)$ ，且假设 σ^2 与 X 无关。

进行一元线性回归主要讨论如下问题：

- (1) 利用样本数据对参数 β_0 ， β_1 和 σ^2 进行点估计，得到经验回归方程；
- (2) 检验模型的拟合程度，验证 Y 与 X 之间的线性相关的确存在，而不是由于抽样的随机性导致的；
- (3) 利用求得的经验回归方程，通过 X 对 Y 进行预测或控制。

8.2.1 简单回归方程的求解

我们希望根据观测值 $(x_i, y_i), i=1, 2, \dots, n$ 估计出简单回归方程中的待定系数 β_0 和 β_1 ，它们使得回归方程对应的响应变量的误差达到最小，该方法即为最小二乘法。

也就是求解 β_0 和 β_1 ，使得

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (8-4)$$

达到最小。

设 \bar{x} 和 \bar{y} 分别为预测变量和响应变量的样本均值； s_x 和 s_y 分别为预测变量和响应变量的样本标准差； $Cor(Y, X)$ 为样本相关系数，则最小二乘法给出的回归系数的估计值为

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = Cor(Y, X) \frac{s_y}{s_x} \quad (8-5)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (8-6)$$

于是把估计值代入到回归方程 8-3，得到拟合回归方程为

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

对于每个样本观测值，我们可以计算

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ，其中 \hat{y}_i 称为相应于 x_i 的拟合值。相应于第 i 个观测值，它的预测误差为

$$e_i = y_i - \hat{y}_i \quad (8-7)$$

SPSS 在输出回归系数的估计值的同时还会给出回归系数估计值的标准误差值，它们的公式为：

$$Var(\hat{\beta}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

这里, $\hat{\sigma}^2$ 为回归方程的误差项 ε 的方差 σ^2 的无偏估计

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} \quad (8-8)$$

注意:

式 8-7 中的预测误差 $e_i, i=1, 2, \dots, n$, 被称为未标准化最小二乘残差, 它们的方差不等的, 我们有时候采用学生化残差, 其计算公式为

$$s_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - p_{ii}}}, i=1, 2, \dots, n,$$

其中, $\hat{\sigma}$ 为 8-8 公式中定义的方差的平方根, p_{ii} 为映射矩阵 ($P = X(X^T X)^{-1} X^T$) 对角线上的第 i 个元素。另外一种形式的残差为标准化的残差, 其计算公式为:

$$z_i = e_i / \sqrt{\sum_{i=1}^n e_i^2}, i=1, 2, \dots, n.$$

另外, 除未标准化残差 e_i 、标准化残差 z_i 和学生化残差 s_i 以外, SPSS 还可以输出删失残差和学生化删失残差。

8.2.2 回归方程拟合程度检验

1. 假设检验

回归方程的检验也就是验证两个变量之间的线性关系的确在统计上显著。一般进行如下的假设检验。

1) 常数项的 t 检验

$$H_0: \hat{\beta}_0 = 0$$

检验统计量为 t 统计量, 其定义为

$$t = \frac{\hat{\beta}_0}{s.e.(\hat{\beta}_0)}$$

其中, $s.e.(\hat{\beta}_0)$ 为 $\hat{\beta}_0$ (常数项的估计值) 的标准误差。即 t 统计量为常数项的估计值和其标准误差的比值。SPSS 回归分析的系数表中会给出回归方程常数项的估计

值、标准误差、t 统计量以及相应的显著性值。

2) 回归系数显著性的 t 检验

$$H_0: \hat{\beta}_1 = 0$$

检验统计量为 t 统计量，其定义为

$$t = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

其中， $s.e.(\hat{\beta}_1)$ 为 $\hat{\beta}_1$ （预测变量 X 的回归系数的估计值）的标准误差。即 t 统计量为参数的估计值和其标准误差的比值。SPSS 回归分析的系数表中会给出回归参数的估计值、标准误差、t 统计量以及相应的显著性值。

3) 相关系数显著性的 t 检验

$$H_0: \hat{\rho} = 0$$

该假设检验用于检验变量 X 和变量 Y 的相关系数是否等于 0。SPSS 在给出两变量 Pearson 相关系数时，可以进行此项检验。

检验统计量为：

$$t_1 = \frac{Cor(Y, X)\sqrt{n-2}}{\sqrt{1-[Cor(Y, X)]^2}}$$

2. 决定系数 R^2

我们把拟合值和真实值的差值的平方和称为残差平方和，记为 SSE；把由于采用拟合回归直线后预测值较采用响应变量均值提高的部分的平方和称为回归平方和，记为 SSR；真实值和响应变量均值的平方和称为总平方和，记为 SST。这里：

$$SST = \sum (y_i - \bar{y})^2 \quad (\text{总平方和})$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \quad (\text{回归平方和})$$

$$SSE = \sum (y_i - \hat{y}_i)^2 \quad (\text{残差平方和})$$

三者之间的关系为：

$$SST = SSR + SSE,$$

定义统计量 $R^2 = SSR / SST = 1 - \frac{SSE}{SST}$ ，称为回归方程的决定系数。由于 $SSR \leq SST$ ，所以 $0 \leq R^2 \leq 1$ 。决定系数的大小反映了回归方程能够解释的响应变量总的变差的比例，其值越大，回归方程的拟合程度越高。但是， R^2 到底多大拟合直线才算满意呢？不同的问题其答案不同，一般而言，在 0.6 以上即可以接受回归直线。

一般情况下，随着预测变量个数的增大，决定系数的值也变大，因此在多重回归分析中，需要反映回归方程中预测变量的个数，即引入了调整的决定系数。请参见第 8.3 节。

3. 回归模型的显著性的 F 检验

总平方和 SST 反映因变量 Y 的波动程度或者不确定性，在建立了 Y 对 X 的回归方程后，总平方和 SST 分解成回归平方和 SSR 与参差平方和 SSE 两部分。其中 SSR 是由回归方程确定的，SSE 是不能由自变量 X 解释的波动，是由 X 之外的未加控制的因素引起的。这样，SST 中能够由自变量解释的部分为 SSR，不能由自变量解释的部分为 SSE。这样回归平方和越大，回归的效果越好，据此构造 F 检验统计量如下

$$F = \frac{SSR / 1}{SST / (n - 2)}$$

SPSS 在回归输出结果的 ANOVA 表中给出 SSR, SSE, SST 和 F 统计量的取值，同时给出 F 值的显著性值（即 p 值）。

注意：对于一元线性回归，回归系数显著性的 t 检验，回归模型的显著性的 F 检验，相关系数显著性的 t 检验的检验结果是完全等价的。其实，可以证明，回归系数显著性的 t 检验与相关系数显著性的 t 检验是完全相等的，而 F 统计量则为这两个 t 统计量的平方。因此，一元线性回归实际上只需要做其中的一种检验即可。然而对于多元线性回归，这三种检验所考虑的问题有所不同，因而并不等价。

8.2.3 用回归方程预测

在一定范围内，对任意给定的预测变量取值 x_0 ，可以利用求得的拟合回归方程进行预测。其预测值为：

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

该预测值的 $(1-\alpha)100\%$ 置信区间为:

$$(\hat{\mu}_0 - t_{n-2, \alpha/2} \times s.e.(\hat{\mu}_0), \hat{\mu}_0 + t_{n-2, \alpha/2} \times s.e.(\hat{\mu}_0))$$

其中 $s.e.(\hat{\mu}_0)$ 为预测值 μ_0 的标准误差, 其估计值为:

$$s.e.(\hat{\mu}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

在 SPSS 回归分析的“保存”选项中, 可以选择保存预测值和其相应的 $(1-\alpha)100\%$ 置信区间。

同时, SPSS 可以提供标准化的预测值和调整的预测值, 其计算公式分别为

$$ZPred_i = \frac{\hat{y}_i - \bar{y}}{s.d.(\hat{y})}$$

$$AdjPred_i = y_i - \frac{e_i}{1 - p_{ii}}$$

8.2.4 简单线性回归举例

一家计算机服务公司需要了解其用电话进行客户服务修复的计算机零部件的个数和其电话用的时间的关系。经过相关分析, 认为二者之间有显著的线性关系。下面我们利用线性回归找到这两个变量之间的数量关系。

在 SPSS 中打开数据文件 ComputerRepair.sav, 变量 Units 记录了修复的零部件的个数; 变量 Minuts 记录了服务所占用的电话时间。

选择【分析】→【回归】→【线性】, 得到如图 8-2 所示的一元线性回归对话框。把 Units 选入到自变量框中; 把 Minuts 选入到因变量框中。其他选项保留默认值。

以上操作可以通过下列语法命令来完成。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 8\ComputerRepair.sav' .
DATASET NAME myData WINDOW=FRONT.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Minuts
  /METHOD=ENTER Units.
```



图 8-2 一元线性回归

结果浏览器中的结果如表 8-1 和表 8-2 所示。

表 8-1 回归系数及其检验

模型	非标准化系数		标准系数	t	Sig.
	B	标准误差	试用版		
1 (常量)	4.162	3.355		1.240	.239
Units	15.509	.505	.994	30.712	.000

a. 因变量: Minutes

表 8-1 给出了线性回归模型的参数估计。SPSS 除了给出非标准化系数，即公式 8-5 和 8-6 中的常规最小二乘估计值以外，还给出标准化预测变量和标准化响应变量后的回归系数。

这里标准化的预测变量和响应变量分别为

$$x_i^* = \frac{x_i - \bar{x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, i = 1, 2, \dots, n. \quad \text{和} \quad y_i^* = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, i = 1, 2, \dots, n.$$

t 列为相应的估计值的 t 检验的 t 统计量的值，Sig. 列为相应的显著性值 (p 值)，Units 的显著性值为 .000，小于 0.05，因此该系数显著区别于零。常量的显著性值大于 0.05，即该项不显著。

表 8-2 模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.994 ^a	.987	.986	5.392

a. 预测变量: (常量), Units。

表 8-2 的模型汇总给出了线性回归的决定系数, $R^2=0.987$, 说明该线性模型可以解释自变量 98.7% 的变差, 拟合效果较好。

表 8-3 模型拟合优度检验

Anova ^b						
模型		平方和	df	均方	F	Sig.
1	回归	27419.509	1	27419.509	943.201	.000 ^a
	残差	348.848	12	29.071		
	总计	27768.357	13			

a. 预测变量: (常量), Units。

b. 因变量: Minutes

表 8-3 的模型拟合优度检验 Anova 表中的 F 检验的显著性值小于 0.05, 表明一元线性回归模型显著。

8.3 多元线性回归

实际应用中, 很多情况要用到多个预测变量才能更好地描述变量间的关系, 如果这些预测变量在预测方程中的系数为线性, 那么回归方程称为多元线性回归方程。就方法的实质来说, 处理多个预测变量的方法与处理一个预测变量的方法基本相同, 只是多元线性回归的方法复杂些, 计算量也大得多, 一般都用计算机进行处理。

8.3.1 多元线性回归方程简介

1. 多元线性回归的模型

设因变量 Y 与自变量 X_1, X_2, \dots, X_p 之间有下列关系式

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (8-9)$$

其中 $\beta_0, \beta_1, \dots, \beta_p$ 为常数, 有时称为 (偏) 回归系数, ε 为随机误差。

它假设对于一定范围内的任何 X_1, X_2, \dots, X_p 的取值, 多元线性回归方程提供了 X 和 Y 的线性关系的近似描述。

设 x_{ij} 是自变量 X_i 的第 j 个观测值, y_j 是因变量 Y 的第 j 个值, 代入公式 8-8 中得到多元线性回归模型的数据结构形式为:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_p x_{p1} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \dots + \beta_p x_{p2} + \varepsilon_2 \\ \dots \quad \dots \quad \dots \quad \dots \\ y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \dots + \beta_p x_{pn} + \varepsilon_n \\ \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \sim N(0, \sigma^2), \text{ 且各个 } \varepsilon_i \text{ 相互独立} \end{cases} \quad (8-10)$$

我们称公式 (8-9) 或公式 (8-10) 为 p 元正态线性回归模型, 其中 $\beta_0, \beta_1, \dots, \beta_p$ 及 σ^2 都是未知待估计的参数, 对多元线性回归模型, 需讨论的问题与简单线性回归相同。

8.3.2 多元线性回归方程的显著性检验

与一元的情形一样, 上面的讨论是在响应变量 Y 与预测变量 X_1, X_2, \dots, X_p 之间呈现线性相关的前提下进行的, 所求的经验方程是否有显著意义, 还需对 Y 与 $X_i, i=1, 2, \dots, p$ 间是否存在线性相关关系作显著性假设检验, 与一元类似, 回归方程是否有显著意义, 需要对回归参数 $\beta_0, \beta_1, \dots, \beta_p$ 进行检验。

1. 检验每个回归系数是否显著

$$H_0: \beta_j = 0$$

这里和一元线性回归的检验一样, 检验统计量为 t 统计量。

2. 检验所有回归系数都不显著

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

这里的检验统计量为

$$F = \frac{SSR / p}{SSE / (n - p - 1)} = \frac{MSR}{MSE}, \text{ 或者记为}$$

$$F = \frac{\text{回归平方和} / p}{\text{残差平方和} / (n - p - 1)} = \frac{\text{回归均方}}{\text{残差均方}}$$

注意: 1. F 检验的 H_0 被拒绝, 并不能说明所有的自变量都对因变量 Y 有显著影响, 我们希望从回归方程中剔除那些统计上不显著的自变量, 重新建立更为简单的线性回归方程, 这就需要对每个回归系数做显著性检验。

2. 即使检验 1 中所有的回归系数单独检验统计上都不显著, 而 F 检验有可能显著, 这时我们不能说模型不显著。这时候, 尤其需要仔细对数据进行分析, 可能分析的数据有问题, 譬如共线性等。

SPSS 输出结果的 ANOVA 表将进行该项检验。

但在实用中, 多元回归中剔除变量的问题比上例我们做的讨论要复杂得多, 因为有些变量单个讨论时, 对因变量的作用很小, 但它与某些自变量联合起来, 共同对因变量的作用却很大, 因此在剔除变量时, 还应考虑变量交互作用对 y 的影响。此外, 关于多元性回归的预测和控制问题, 类似一元不再赘述。

8.3.3 应用举例

本章的数据文件 performance.sav 记录了一项企业心理学研究的数据。它调查了一个大型金融机构的雇员, 记录了他们和主管的交互情况的评价和对主管的总的满意情况。我们希望该调查来了解主管的某些特征和对他们的总的满意情况的相互关系。

打开数据文件 performance.sav, 选择【分析】→【回归】→【线性】, 得到如图 8-3 所示的多元线性回归对话框。把变量 Y 选入到因变量框中, 把变量 $X1$ 到 $X6$ 选入到自变量框中, 其他选项保留默认值。单击【确定】按钮。



图 8-3 多元线性回归

以上操作可以通过下列语法命令来完成。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = 'C:\SPSSIntro\Chapter 8\performance.sav' .
DATASET NAME myData WINDOW=FRONT.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Y
  /METHOD=ENTER X1 X2 X3 X4 X5 X6.
```

在结果查看器中得到表 8-3 中的三张表格，它们分别为模型汇总、ANOVA 和系数。

其中的“模型汇总”表输出 R 、 R 方和调整 R 方。 R 方统计量 0.733 表明该线性模型可以解释自变量 73.3% 的变差。一般而言，随着自变量个数的增多，不管增加的自变量是否和因变量的关系密切与否， R 方都会增大；调整的 R 方是根据回归方程中的参数的个数进行调整的 R 方，它对参数的增多进行惩罚，调整 R 方没有直观的解释意义，它的定义为：

$$R_{\text{调整}}^2 = 1 - \frac{SSE / (n - p - 1)}{SST / (n - 1)} = 1 - \frac{n - 1}{n - p - 1} (1 - R^2)$$

“ANOVA”表为模型显著性 F 检验的结果，它输出多元线性回归模型的平方和、自由度、均方、F 值和相应的显著性值。平方和列为回归平方和 SSR、残差平方和 SSE、总平方和 SST，均方列为 MSR 和 MSE。这里显著性值为 .000，小于 0.05，即该回归模型显著。

“系数”表为模型的未标准化回归系数的估计值、标准化变量后的回归系数的估计值、t 检验的 t 统计量值和相应的显著性值。非标准化系数列记录了最小二乘法在原始数据上的线性回归系数的估计值。

“标准系数”列记录了最小二乘法在标准化后的数据上的线性回归系数的估计值。

标准化预测变量和标准化响应变量的公式分别为：

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}, i=1,2,\dots,n; j=1,2,\dots,p \text{ 和 } y_i^* = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, i=1,2,\dots,n$$

“t”列记录了各回归系数 t 检验的 t 统计量，而 Sig.列记录了相应的显著性值。这里，只有 X1 和 X3 的显著性值小于 0.1，注意到回归方程的常数项也不显著。然而，大部分情况下不显著的预测变量都要从回归方程中移除，而回归常数代表了响应变量的基本水平，不管显著与否，大部分情况都保留在回归方程中。因此，我们可以仅仅考虑 Y 和 X1、X3 之间的关系而忽略其他预测变量。

表 8-3 多元线性回归结果

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.856 ^a	.733	.663	7.068

a. 预测变量: (常量), X6, X1, X5, X2, X3, X4。

Anova^b

模型		平方和	df	均方	F	Sig.
1	回归	3147.966	6	524.661	10.502	.000 ^a
	残差	1149.000	23	49.957		
	总计	4296.967	29			

a. 预测变量: (常量), X6, X1, X5, X2, X3, X4。

b. 因变量: Y

系数^a

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	10.787	11.589		.931	.362
	X1	.613	.161	.671	3.809	.001
	X2	-.073	.136	-.073	-.538	.596
	X3	.320	.169	.309	1.901	.070
	X4	.082	.221	.070	.369	.715
	X5	.038	.147	.031	.261	.796
	X6	-.217	.178	-.183	-1.218	.236

a. 因变量: Y

把 Y 作为因变量，X1 和 X3 作为自变量，重复进行回归分析。操作界面如图 8-4 所示。



图 8-4 重复进行多元线性回归

以上操作可以通过以下语法命令完成。

```
REGRESSION  
  /MISSING LISTWISE  
  /STATISTICS COEFF OUTS R ANOVA  
  /CRITERIA=PIN(.05) POUT(.10)  
  /NOORIGIN  
  /DEPENDENT Y  
  /METHOD=ENTER X1 X3.
```

结果查看器中得到回归分析的结果如表 8-4 所示。

从表 8-4 中的模型汇总知，虽然由于预测变量的个数由 6 个减少为 2 个， R 方略有变小，但是调整的 R 方变大了，根据模型精简的原则。我们倾向于采用新的精炼的模型。

在“系数”表中， $X3$ 对应的显著性值大于 0.1，但是根据问题的实际背景，我们认为该预测变量和总的满意度 Y 是有关系的，因此我们在回归模型中保留了该预测变量。

表 8-4 多元回归分析结果

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.841 ^a	.708	.686	6.817

a. 预测变量: (常量), 有学习新东西的机会, 处理员工抱怨。

Anova^b

模型		平方和	df	均方	F	Sig.
1	回归	3042.318	2	1521.159	32.735	.000 ^a
	残差	1254.649	27	46.468		
	总计	4296.967	29			

a. 预测变量: (常量), 有学习新东西的机会, 处理员工抱怨。

b. 因变量: 总的满意情况

系数^a

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	9.871	7.061		1.398	.174
	处理员工抱怨	.644	.118	.704	5.432	.000
	有学习新东西的机会	.211	.134	.204	1.571	.128

a. 因变量: 总的满意情况

8.4 线性回归的诊断和线性回归过程中的其他选项

不论是一元线性回归方程还是多元线性回归方程, 在应用时都要求满足一定的前提条件。前几节中的最小二乘法 and 所有假设检验等都是基于这些前提条件的。对于最小二乘法, 如果这些条件有较小的违背, 对回归分析模型的结果影响不大; 但是, 如果严重违背这些条件的话, 有可能得到严重扭曲的结论。对这些前提条件进行验证是回归分析中必不可少的一环。

8.4.1 回归分析的前提条件

回归分析的前提条件可以归纳为下列几条。

- 响应变量和预测变量之间的关系必须为线性关系。可以通过考察散点图来进行验证。
- 线性回归模型的误差变量是服从相互独立的、分布相同的正态分布

$N(0, \sigma^2)$, 即:

- 1) 误差变量 $\varepsilon_i, i=1, 2, \dots, n$ 为正态分布;
 - 2) $\varepsilon_i, i=1, 2, \dots, n$ 的均值为 0;
 - 3) $\varepsilon_i, i=1, 2, \dots, n$ 有相同的方差 σ^2 ;
 - 4) 误差变量 $\varepsilon_i, i=1, 2, \dots, n$ 是相互独立的。
- 预测变量的取值没有测量误差。该项较难验证, 一般而言, 如果测量误差相对随机误差不是太大, 那么测量误差的影响可以忽略。
 - 预测变量相互之间线性无关。该前提条件可以保证最小二乘的解是唯一的, 如果违反该条件, 则出现共线性问题。
 - 所有的观测值的在分析中的作用是相同的。

8.4.2 回归分析前提条件的检验

1. 线性相关性的检验

只有变量间真实存在线性相关关系时, 才可以应用线性回归模型来建模。仅仅观察两个变量之间的相关系数有时候是不够的。例如 Anscombe 构造的四组数据, 如表 8-5 所示, 它们的相关系数都是 0.816, 并且统计检验是显著的; 用它们进行线性回归, 得到的四组拟合直线的斜率也完全一致。通过观察图 8-5 中的散点图知道, 只有图 8-5 (a) 才有线性关系, 图 8-5 (b) 中的关系是非线性的, 图 8-5 (c) 中有一个离群点, 导致了回归直线产生系统性的偏差; 图 8-5 (d) 中的数据由于一个离群点的存在, 导致了拟合的回归直线大大偏离了实际数据的情况。

表 8-5 Anscombe 构造的四组数据

y1	x1	y2	x2	y3	x3	y4	x4
8.04	10.00	9.14	10.00	7.46	10.00	6.58	8.00
6.95	8.00	8.14	8.00	6.77	8.00	5.76	8.00
7.58	13.00	8.74	13.00	12.74	13.00	7.71	8.00
8.81	9.00	8.77	9.00	7.11	9.00	8.84	8.00
8.33	11.00	9.26	11.00	7.81	11.00	8.47	8.00
9.96	14.00	8.10	14.00	8.84	14.00	7.04	8.00
7.24	6.00	6.13	6.00	6.08	6.00	5.25	8.00
4.26	4.00	3.10	4.00	5.39	4.00	12.50	19.00
10.84	12.00	9.13	12.00	8.15	12.00	5.56	8.00
4.82	7.00	7.26	7.00	6.42	7.00	7.91	8.00
5.68	5.00	4.74	5.00	5.73	5.00	6.89	8.00

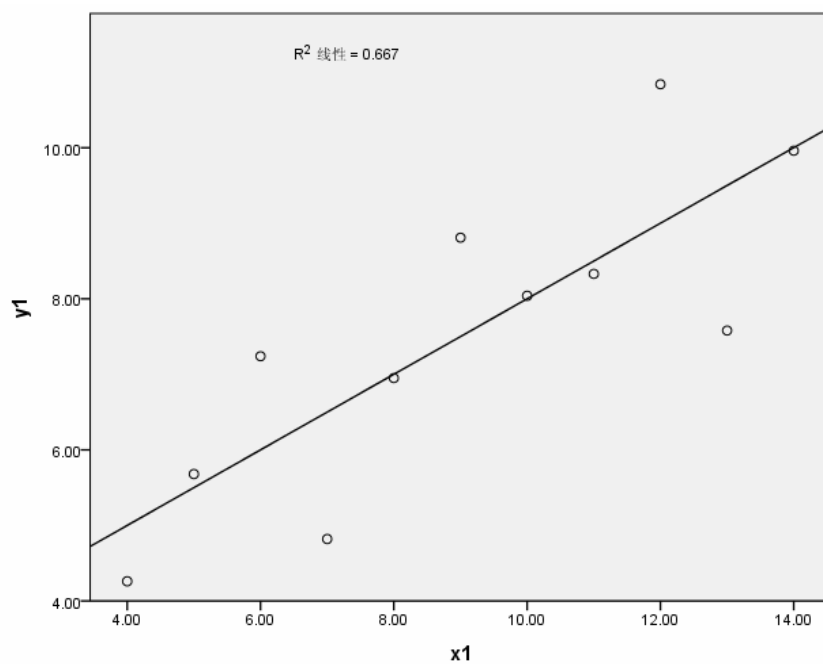


图 8-5(a) X1-Y1 散点图

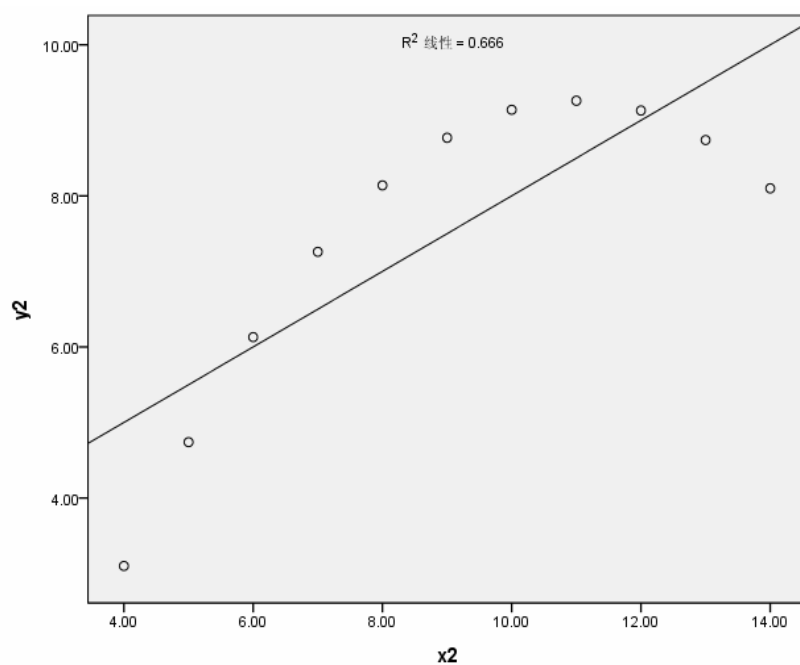
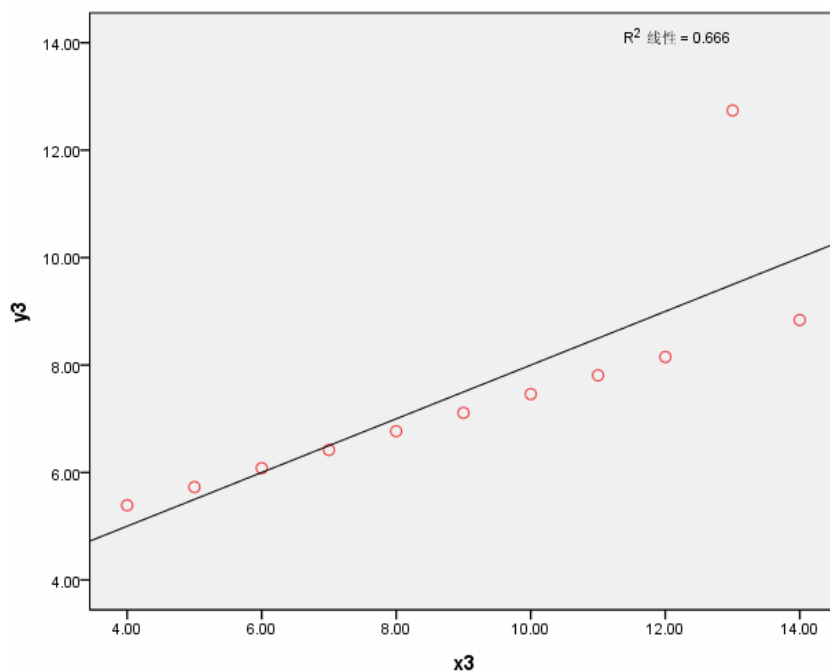
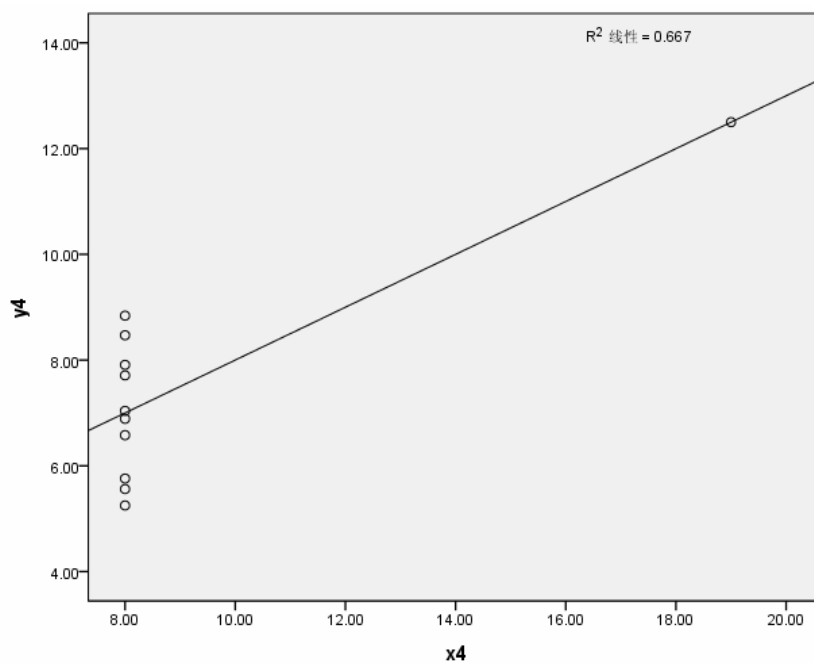


图 8-5(b) X2-Y2 散点图

图 8-5(c) X_3 - Y_3 散点图图 8-5(d) X_4 - Y_4 散点图

2. 图形方法

在统计分析中，应用统计图进行分析是最直观也是很有效率的一种方法。

1) 在进行模型拟合之前

用以表达预测变量和响应变量之间关系的模型应该基于理论基础或者某种有待检验的假定。如果对变量之间的关系(即模型的形式)没有任何先验知识,我们可以绘制散点图,直方图、点图、箱图和茎叶图等来进行数据的探索性分析。另外,可以绘制二维或者三维的散点图来分析变量之间的关系,找出他们相关的模式。

除了以上图形外,还可以进行图形的旋转,绘制动态图形等方法,辅助进行探索性的数据分析工作。

2) 在模型拟合之后

模型拟合完成后,通过统计图来检验线性回归的前提条件是否满足,模型拟合是否满意等。可以完成以下工作:

- 通过统计图形来检查模型的线性和误差的正态性假设;
- 通过统计图来检查离群值,影响点;
- 各变量影响效果的诊断图。

SPSS 线性回归过程中的“绘制(T)”选项,提供各种残差图供用户诊断模型的拟合程度。残差图如图 8-6 所示,用户可以选择绘制响应变量和标准化残差之间、响应变量和标准化预测值之间,等等的残差图。



图 8-6 残差图

这些变量之间的散点图可以帮助验证正态性、线性和方差相等的假设。对于检测离群值、异常观察值和有影响的个案,这些散点图也是有用的。在将它们保存为

新变量之后，在数据编辑器中可以使用预测值、残差和其他诊断以使用自变量构造图。图 8-6 中的这些图是可利用的。

变量列表：左边框中列出因变量(DEPENDNT)及相关的预测变量和残差变量，它们包括：标准化预测值(*ZPRED)、标准化残差(*ZRESID)、剔除残差(*DRESID)、调整的预测值(*ADJPRED)、学生化的残差(*SRESID)以及学生化的已删除残差(*SDRESID)。

散点图：您可以绘制以下各项中的任意两种：响应变量、标准化预测值、标准化残差、剔除残差、调整预测值、Student 化的残差或 Student 化的已删除残差。针对标准化预测值绘制标准化残差，以检查线性关系和等方差性。通过选择“下一张”可以绘制多幅散点图。

产生所有部分图：当根据其余自变量分别对两个变量进行回归时，显示每个自变量残差和因变量残差的散点图。要生成部分图，方程中必须至少有两个自变量。从该图中可以判断各个预测变量和响应变量的线性相关性。

标准化残差图：可以获取标准化残差的直方图和正态概率图（即标准化残差的 P-P 图），将标准化残差的分布与正态分布进行比较，以验证残差的正态性。

3. 离群值和影响点的探测

我们希望得到的回归方程不被一个或者少数几个点影响。如果一个或者少数几个观测值去掉之后，回归方程发生极大的变化，这些观测值被称为影响点。需要注意的是，通过残差散点图很难找到影响点，在许多情况下，影响点的残差不是很大，影响点的残差不是离群值。它们对回归直线的影响却大于其他观测值。但是，移除影响点后，回归直线会发生极大的变化。

我们可以通过寻找响应变量和预测变量中的离群值来寻找影响点。

SPSS 输出一些指标来帮助标识影响点，例如 Cook 距离、杠杆值等。在“回归”对话框中，单击“保存”，选择“Cook 距离(K)”和“杠杆值(G)”，那么各个观测值的 Cook 距离和杠杆值就保存在数据视图中。如图 8-7 所示。

图 8-7 中可以选择保存预测值、残差和其他对于模型诊断有用的统计量。每选择一次将向当前数据文件添加一个或多个新变量。图 8-9 中相关选项的意义如下：

预测值：回归模型对每个个案预测的值

- 未标准化。保存应用未标准化系数模型对因变量的预测值。
- 标准化：每个预测值转换为其标准化形式的转换。即预测值减去均值预测值，得到的差除以预测值的标准差。标准化预测值的均值为 0，标准差为 1。

距离：用以判定各个个案对拟合直线的影响程度

- Cook 距离：在特定个案从回归系数的计算中排除的情况下，所有个案的残差变化幅度的测量。较大的 Cook 距离表明从回归统计量的计算中排除个案之后，系数会发生根本变化。



图 8-7 保存预测值、残差、距离（探测影响点的指标）和影响统计量

- 杠杆值：度量某个点对回归拟合的影响。集中的杠杆值范围为从 0（对拟合无影响）到 $(N-1)/N$ 。一般情况如果杠杆值大于 0.06，就要引起注意。

影响统计量：由于排除了特定个案而导致的回归系数(DfBeta)和预测值(DfFit)的变化。标准化 DfBeta 和 DfFit 值也可与协方差比率一起使用。

- **DfBeta(B)**: beta 值的差分是由于排除了某个特定个案而导致的回归系数的改变。为模型中的每一项（包括常数项）均计算一个值。
- **标准化 DfBeta**: beta 值的标准化差分。由于排除了某个特定个案而导致的回归系数的改变。您可能想要检查除以 N 的平方根之后绝对值大于 2 的个案，其中 N 是个案数。为模型中的每一项（包括常数项）均计算一个值。
- **DfFit (F)**: 拟合值的差分是由于排除了某个特定个案而产生的预测变量的更改。
- **标准化 DfFit**: 拟合值的标准化差分。由于排除了某个特定个案而导致的预测值的改变。您可能想要检查绝对值大于 p/N 的平方根的 2 倍的标准值，其中 p 是模型中的参数个数， N 是个案数。
- **协方差比率**: 从回归系数计算中排除特定个案的协方差矩阵的行列式与包含所有个案的协方差矩阵的行列式的比率。如果比率接近 1，则说明被排除的个案不能显著改变协方差矩阵。

通过分析 Cook 距离、杠杆值，可以识别影响点。以表 8-5 Anscombe 构造的四组数据中的第三组数据 (X_3, Y_3) 为例，我们从图 8-5 散点图知道，该组数据存在影响点。我们绘制该组数据的 Cook 距离和杠杆值的点图，如图 8-8 和图 8-9 所示。

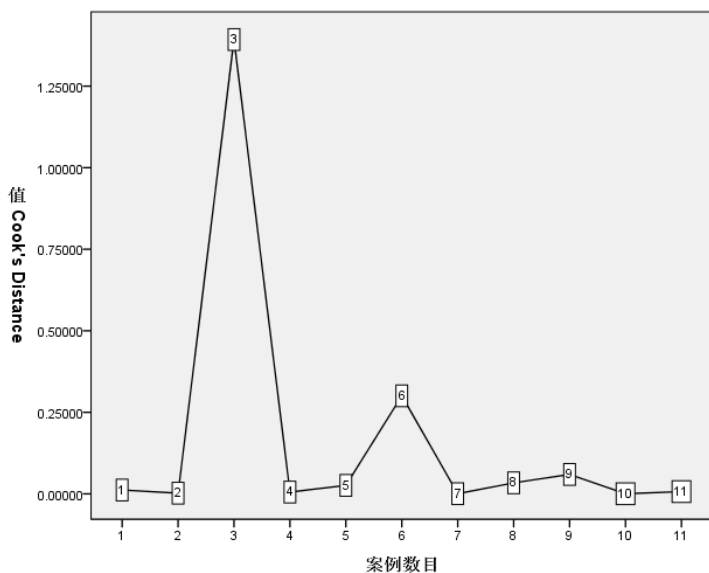


图 8-8 Cook 距离

从 Cook 距离看出，第三个观测和第六个观测为影响点。

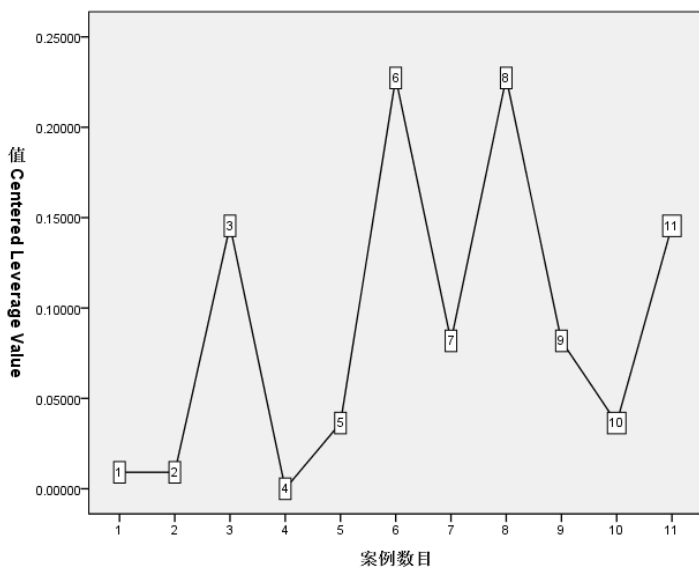


图 8-9 杠杆值

从图 8-9 知，第三个、第六个和第九个观测值为影响点。

结合图 8-8、图 8-9 以及 X_3 和 Y_3 的散点图，我们可以判断第三个观测值为影响点，可以考虑移除掉该观测值后进行回归分析建模。从图 8-10 看出，移除第三个观测值后对回归分析的影响最大。

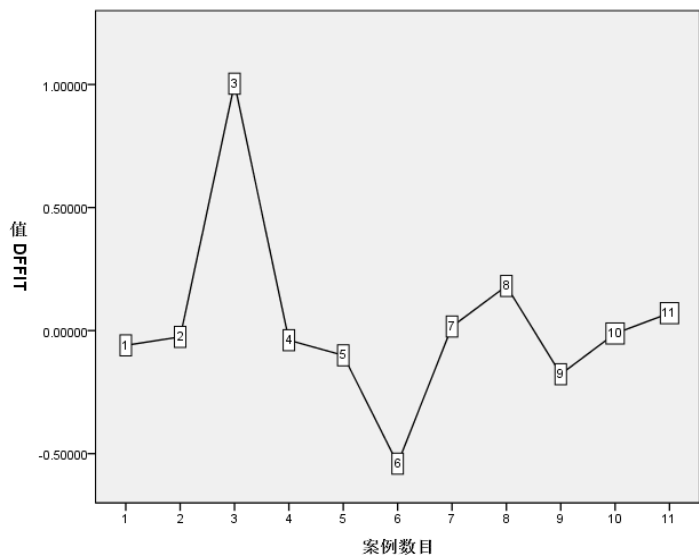


图 8-10 Dfit 值

4. 预测变量共线性的诊断

在 SPSS 的回归分析对话框中，单击“统计量”按钮，如图 8-11 所示。

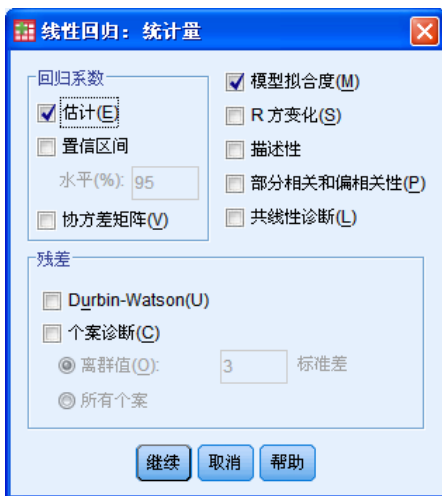


图 8-11 统计量选择—共线性诊断

共线性诊断：共线性（或者多重共线性）是非理想情况，此时一个自变量是其他自变量的线性函数。显示已标度和未中心化交叉积矩阵的特征值、条件指数以及方差-分解比例，以及个别变量的方差膨胀因子(VIF)和容差。

残差：显示残差的序列相关系数的 Durbin-Watson 检验，以及满足选择条件（n 倍标准差以外的离群值）的个案诊断。

8.4.3 线性回归的其他选项

1. 预测变量进入模型的方法

自变量进入模型的方式，如图 8-12 所示。



图 8-12 自变量进入模型的方式

在图 8-12 的“方法(M)”框中允许用户指定自变量将如何进入到分析中。通过使用不同的方法,可以从相同的变量组构造多个回归模型。变量选择过程用下列方法:

- 进入: 一个块中的所有变量(位于一张输入框中的变量)在一个步骤中输入。
- 逐步: 在每一步,不在方程中的具有 F 的概率最小的自变量被选入(如果该概率足够小)。对于已在回归方程中的变量,如果它们的 F 概率变得足够大,则移去这些变量。如果不再有变量符合包含或移去的条件,则该方法终止。
- 删除: 在单步中移去一个块中的所有变量。
- 向后去除: 在该过程中将所有变量输入到方程中,然后按顺序移去。会考虑将与因变量之间的部分相关性最小的变量第一个移去。如果它满足消除条件,则将其移去。移去第一个变量之后,会考虑下一个将方程的剩余变量中具有最小的部分相关性的变量移去。直到方程中没有满足消除条件的变量,过程才结束。
- 向前选择: 一个逐步变量选择过程,在该过程中将变量顺序输入到模型中。第一个考虑要选入到方程中的变量是与因变量之间具有最大的正或负的相关性的变量。只要在该变量满足选入条件时才将它选入到方程中。选入了第一个变量之后,接下来考虑不在方程中的具有最大的部分相关性的自变量。当无满足选入条件的变量时,过程结束。

输出中的显著性值基于与单个模型的拟合。所以,当使用逐步推进方法(逐步式、向前或向后)时,显著性值通常无效。

无论指定什么进入方法,所有变量都必须符合容差条件才能进入方程。缺省的容差水平为 0.0001。另外,如果某个变量会导致另一已在模型中的变量的容差下降到容差条件以下,则该变量不进入方程。

所有被选自变量将被添加到单个回归模型中。不过,您可以为不同的变量子集指定不同的进入方法。例如,您可以使用逐步式选择将一个变量块输入到回归模型中,而使用向前选择输入第二个变量块。

分析中包含由选择规则定义的个案。例如,如果选择变量,选择等于,并为该值键入 5,则只有那些选定变量值等于 5 的个案才会包含在分析中。字符串值也是允许的。

2. 线性回归步进方法的选项

在图 8-13 中，可以设置步进法变量进入模型或者变量被剔除出模型的判断标准。单击图 8-12 中的“选项”按钮，得到如图 8-13 所示的对话框。

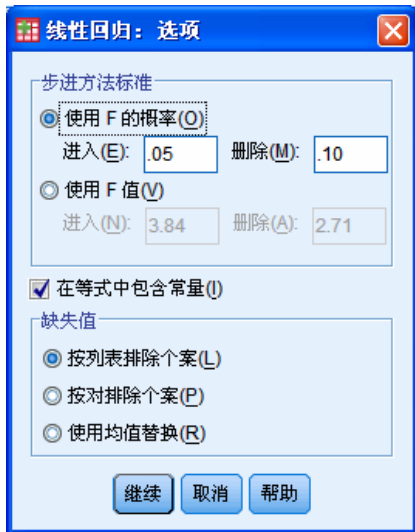


图 8-13 步进法进入标准

步进方法标准：这些选项在已指定向前、向后或逐步式变量选择法的情况下适用。变量可以进入到模型中，或者从模型中移去，这取决于 F 值的显著性（概率）或者 F 值本身。

- **使用 F 的概率：**如果变量的 F 值的显著性水平小于“输入”值，则将该变量选入到模型中，如果该显著性水平大于“剔除”值，则将该变量从模型中移去。“输入”值必须小于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请增加“输入”值。要将更多的变量从模型中移去，请降低“剔除”值。
- **使用 F 的值：**如果变量的 F 值大于“输入”值，则该变量输入模型，如果 F 值小于“剔除”值，则该变量从模型中移去。“输入”值必须大于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请降低“输入”值。要将更多的变量从模型中移去，请增大“剔除”值。

在等式中包含常量：默认情况下，回归模型包含常数项。取消选择此选项可强制使回归直线通过原点，实际上很少这样做。某些通过原点的回归结果无法与包含常数的回归结果相比较。例如，不能以通常的方式解释 R^2 。

8.5 小结

本章主要介绍了回归分析的基本概念。从简单线性回归的概念入手，介绍了回归方程、回归方程的拟合程度检验、应用回归方程进行预测等方法。多元线性回归和简单线性回归十分类似，应用回归分析需要检验回归分析的前提条件。另外，还需要对回归方程的拟合程度进行分析和检验。SPSS 线性回归分析过程提供了丰富的选项，它可以根据预测变量与响应变量的相关程度来选择预测变量。

思考与练习

1. 数据文件 world95.sav 记录了 1995 年统计的各个国家的生育率(fertility)和妇女的平均预期寿命(lifeexpf)等数据。

- 1) 探索性分析这两个变量，探索两个变量中是否存在异常点。
- 2) 做出这两个变量的散点图，建立两个变量的线性回归模型，判断得到的模型的合理性。
- 3) 利用生育率来预测妇女的预期寿命。并设置相关选项，以进一步检验关于线性回归的假定条件。判断该数据是否满足线性回归的假定条件。
- 4) 并进行回归诊断，对模型的系数进行解释。从输出结果，判断妇女多要一个小孩对她的寿命的影响情况。

2. 数据文件 FoodConsum.sav 记录了我国 31 个省市自治区的人均食品支出与人均收入的有关数据。请分析人均食品支出与人均收入的依存关系。

3. 下面哪些指标能够给出个案对回归影响大小的信息

- A) COOK 距离
- B) R 方
- C) 变化的 R 方
- D) Leverage 值

4. 进行线性回归，需要对回归的条件进行验证，哪些条件是不需要验证的：

- A) 因变量和自变量之间具有因果关系
- B) 残差具有方差齐性
- C) 残差之间不相关
- D) 自变量服从正态分布

5. 在一元回归情况下，以下论断正确的是：

- A) 回归方程的显著性检验和斜率的显著性检验是等价的
- B) R 方和变化的 R 方等价

- C) 回归方程的常数项可以忽略
- D) 以上论断都不正确

参考文献

1. 卢淑华. 社会统计学 (第三版). 北京: 北京大学出版社, 2005。
2. 吴喜之. 非参数统计 (第二版). 北京: 中国统计出版社, 2006。
3. Michael Sullivan, III, Statistics, NJ: Prentice Hall, 2003。
4. 何晓群. 现代统计分析方法与应用 (第二版). 北京: 中国人民大学出版社, 2007。

本章学习目标:

- 掌握方差分析的基本思想;
- 了解方差分析和比较均值的异同;
- 掌握单因素方差分析的应用条件、方法和结果的解释;
- 掌握多因素方差分析的应用条件、方法和结果的解释;
- 掌握协方差分析的应用条件、方法和结果的解释。

在第 5 章均值的比较中, T 检验应用于研究单样本均值的比较和两个样本均值的比较。在生产活动和科学研究中经常会遇到比较三个或者三个以上样本均值的差异问题。这时,采用的统计方法称为方差分析,简称 ANOVA(ANalysis Of Variance)。例如某机构对当前民众的生活状况进行调查,根据被调查者的回答把居民对待生活的态度分为三类:认为生活丰富多彩、生活平平常常和生活乏味三类,它们想知道人们对待生活的态度是否和他们受教育的情况有关系,即这三类人是否在受教育程度上有显著的区别。又例如,某公司在不同地区采用了三种不同广告形式的促销活动,它们希望对广告形式的效果进行分析,同时还想了解是否不同广告形式是否在不同地区的效果是不同的。这两个例子所涉及的问题分别为单因素的方差分析和两因素的方差分析问题。

方差分析由 R.A. Fisher 创立,它的思想是将总的方差分解为由于随机抽样引起的差异(个体间差异)和由于研究因素所造成的差异两部分,然后比较这两部分差异在总方差中所占的比重。因此,方差分析是利用分析方差的分解来实现对总体均值的比较的。SPSS 的方差分析除了分析总体均值间的差异外,还能够指出哪些总体均值之间存在着显著差异,哪些总体均值之间的差异统计上不显著。

本章 9.1 节引入方差分析中用到的一些术语,9.2 节到 9.4 节分别介绍单因素的方差分析、多因素的方差分析和协方差分析。

9.1 方差分析的术语与前提

方差分析最早是起源于分析试验数据，而试验中则涉及影响试验结果的许多因素。试验中的实验结果是需要分析的变量，称为响应变量，或者因变量。方差分析的因变量必须为尺度类型的数据（即连续数据）。影响试验结果的因素即为影响响应变量的变量，称为自变量或者因子。根据试验中这些因素的处理方式，因素可以分为控制因素、随机因素和协变量。因子的不同取值称为因子的不同水平。控制因素一般要求为分类变量，而协变量要求为尺度数据。

- 控制因素：它是试验中可以控制的影响试验结果的因素，因素的不同水平会导致不同的试验结果。
- 不可控因素：因素的水平与试验结果的关系是随机的，即不确定因素，但是不同于随机因素，可以理解为非研究关心的因素或非处理因素。
- 随机因素：因素与试验结果的关系是随机的，其水平也是随机出现的。
- 处理：在试验中，控制因素的一个水平或者几个控制因素的某一水平组合称为一个处理。

方差分析就是研究不同的控制因素以及控制因素的不同水平（ >2 ）对试验结果影响有无差异的一种统计分析方法。根据控制因素的个数，方差分析可以分为单因素方差分析和多因素方差分析。根据响应变量的个数，方差分析可以分为单变量方差分析和多变量方差分析。第 5 章中介绍的 T 检验可以看做是单因素双水平的方差分析问题，它是方差分析的一种特殊情况。

方差分析的自变量是“因子”或者“因素”，它是分类变量；其因变量则为尺度变量，需要满足以下两个基本前提条件：

- 每个处理的因变量为正态分布（正态性）；
- 每个处理的因变量具有相同的方差（方差齐性）。

9.2 单因素的方差分析

单因素方差分析用于研究一个影响因素对试验结果的影响，它用于比较两个或者两个以上的总体之间是否有显著的差异。SPSS 的单因素方差分析提供下列分析结果：

- 试验结果在不同组别的统计；
- 检验各个组别方差是否相等；

- 各个组别的概略图（均值图）；
- 配对多重比较；
- 不同组别组合的对比检验；
- 同类子集。

如果对各组比较的总体没有任何先验知识，方差分析检验到各个总体均值间存在显著差异时，SPSS 单因素方差分析提供了两类比较均值的方法：先验对比和两两比较检验。先验对比是在试验开始前进行的检验，而两两比较检验则是在试验结束后进行的。同时 SPSS 提供了分析各个总体趋势的方法。

应用方差分析的因变量需要满足正态性和方差齐性条件。

尽管数据应对称，但方差分析对于偏离正态性是稳健的。各组应来自方差相等的总体。为了检验这种假设，请使用 Levene 的方差齐性检验。

销售经理想了解新员工培训的最佳方式。目前有三种新员工培训方式：为期 1 天的培训、为期 2 天的培训和为期 3 天培训。现在需要比较用这三种方式培训员工的效果，分析这三种培训方式培训员工的效果是否有显著的差异，如果有差异，哪种培训方式最佳。

打开数据文件 salesperformance.sav，它包含两个变量，“组”变量记录了培训方式；“得分”是对员工培训效果的评价。

9.2.1 描述性数据分析

在进行方差分析前，首先需要检验方差分析的前提条件是否满足，如果不满足，看偏离是否严重。然后，根据情况决定是采用方差分析还是采用非参数的方法。

我们首先绘制三种培训方式的误差条形图（如图 9-1 所示），直观地检验各个组别的方差是否相等。然后通过 Levene 检验来检验方差是否相等。

从图 9-1 中的误差条形图可知，随着培训天数的增加，培训考试得分的均值越高。并且，培训 3 天的标准误差最小。比较三个误差条，直观看三种培训方式的方差并不相等。在单因素方差分析中，可以选择进行方差的齐性检验，进一步确认方差齐性条件是否满足。

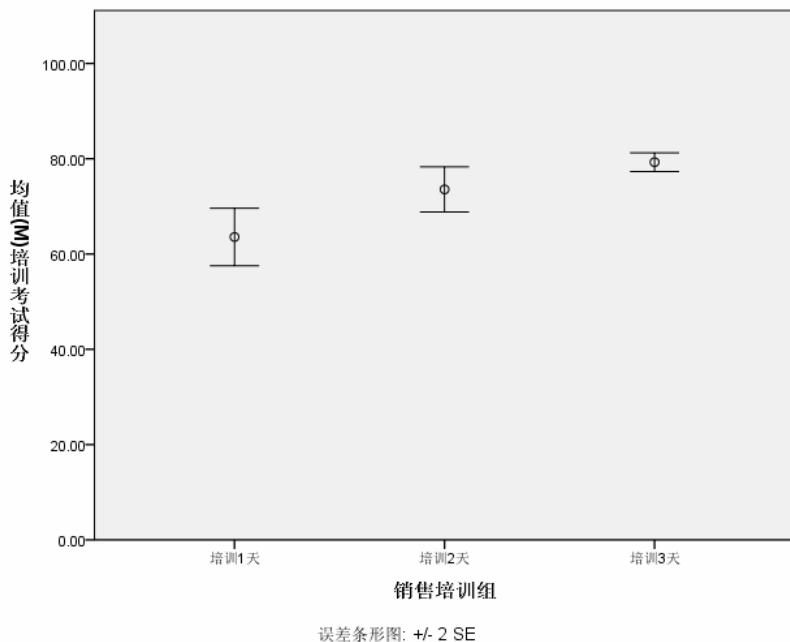


图 9-1 误差条形图

9.2.2 单因素方差分析

打开数据文件 Salesperformance.sav 的数据视图，选择【分析】→【比较均值】→【单因素 ANOVA】，进入“单因素方差分析”窗口，如图 9-2 所示。

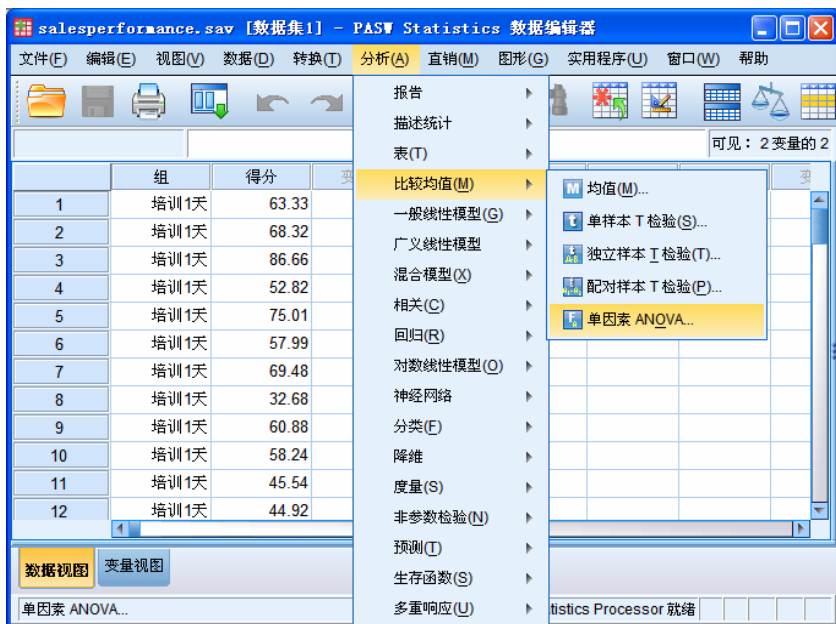


图 9-2 单因素方差分析窗口

单因素方差分析窗口中，把变量“得分”选入“因变量列表(E)”；把“组”选入“因子(F)”框中。如图 9-3 所示。



图 9-3 单因素方差分析对话框

单击“两两比较”按钮，在“假定方差齐性”部分，勾选“Bonferroni(B)”和“Tukey s-b(K)”两项。在“未假定方差齐性”部分，勾选“Tamhane's T2(M)”，如图 9-4 所示。



图 9-4 两两比较

单击【选项】按钮，在【统计量】部分勾选“描述性”和“方差同质性检验”两项，同时勾选“均值图(M)”，如图 9-5 所示。



图 9-5 方差分析选项

单击【继续】按钮，返回上级菜单，单击【确定】按钮。

以上操作可以通过下列语法命令完成。

```
NEW FILE.  
DATASET CLOSE ALL.  
GET FILE = ' C:\SPSSIntro\Chapter 9\salesperformance.sav'.  
DATASET NAME myData WINDOW=FRONT.  
ONEWAY 得分 BY 组  
  /STATISTICS DESCRIPTIVES HOMOGENEITY  
  /MISSING ANALYSIS  
  /POSTHOC=BTUKEY BONFERRONI T2 ALPHA(0.05).
```

在结果查看器中，得到如表 9-1 到 9-4 和图 9-6 所示的结果。

1. 各总体均值之间是否有显著差异

表 9-1 方差分析描述性统计量

描述								
培训考试得分								
	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
培训1天	20	63.5798	13.50858	3.02061	57.2576	69.9020	32.68	86.66
培训2天	20	73.5677	10.60901	2.37225	68.6025	78.5328	47.56	89.65
培训3天	20	79.2792	4.40754	.98556	77.2165	81.3420	71.77	89.69
总数	60	72.1422	12.00312	1.54960	69.0415	75.2430	32.68	89.69

从表 9-1 的描述性统计量结果知，随着培训天数的增加，培训考试得分也随之增加；随着培训天数的增加，培训得分的变化变小，培训 3 天的标准差小于培训 1

天和培训 2 天的标准差。这些差别统计上是否显著呢？

表 9-2 方差齐性检验和 ANOVA 表

方差齐性检验			
培训考试得分			
Levene 统计量	df1	df2	显著性
4.637	2	57	.014

ANOVA					
培训考试得分					
	平方和	df	均方	F	显著性
组间	2525.691	2	1262.846	12.048	.000
组内	5974.724	57	104.820		
总数	8500.415	59			

表 9-2 的“方差齐性检验”部分是对三种培训方式得分的方差是否相等进行检验。这里显著性值为 0.014，小于 0.05，没有理由认为三个组的方差相等。

在比较的各个组别样本量相差不大，并且各组别的分布形态类似的情况下，方差分析对方差不等具有稳健性。该例中，每组的个案数相等，峰度和偏度相等，分布形态类似，因此仍然可以进行方差分析。一般建议，在方差分析之后再运行相应的非参数检验方法来验证方差分析的结果。

“ANOVA”表中“组间平方和”为三种不同培训的均值和总体均值差异的平方和；“组内平方和”为三种培训方式的考试得分和其相应的组考试得分均值差的平方和。

“df”列为自由度。一共有三个组，因素组间自由度为 2；共有 60 个案，三个组，所以组内自由度为 57。

“均方”列为“平方和”列除以相应的自由度，分别称为“组间均方”和“组内均方”。

“F”列为组间均方和组内均方的比值，即

$$F = \frac{2525.691 / 2}{5974.724 / 57} = \frac{1262.846}{104.820} = 12.048$$

相应的显著性值为 0.000，小于 0.05，没有证据说明三种不同的培训方式的效果

是一样的。那么最终应该采用哪种培训方式呢？需要对三种培训方式的两两比较和同类子集进行分析。

2. 均值的两两比较

表 9-3 的多重比较给出了方差相等时的 Bonferroni 两两比较和方差不等时的 Tamhane 两两比较。由于 Levene 检验没有证据说明三种培训方式的方差相等，参照两种不同的两两比较的结果是必要的。在多重比较表中，第一列有三部分，第一部分为采用的多重比较的方法，第二部分为比较的参照（I），第三部分为比较的组别（J）；第二列为比较的两列的均值差（I-J）；第三列到第五列分别为均值差的标准误差、显著性值和 95% 的置信区间。如果均值差在 5% 的显著性水平下显著区别于 0，则在右上方会标识一个（*）。

本例中，Bonferroni 和 Tamhane 多重比较的结果是一致的。即培训 2 天和培训 3 天没有显著的区别，而培训 1 天与另外两种培训都有显著区别。

表 9-3 多重比较

多重比较

因变量: 培训考试得分

(I) 销售培训组		(J) 销售培训组	均值差 (I-J)	标准误	显著性	95% 置信区间	
						下限	上限
Bonferroni	培训1天	培训2天	-9.98789 [*]	3.23759	.009	-17.9740	-2.0018
		培训3天	-15.69947 [*]	3.23759	.000	-23.6856	-7.7134
	培训2天	培训1天	9.98789 [*]	3.23759	.009	2.0018	17.9740
		培训3天	-5.71158	3.23759	.249	-13.6977	2.2745
	培训3天	培训1天	15.69947 [*]	3.23759	.000	7.7134	23.6856
		培训2天	5.71158	3.23759	.249	-2.2745	13.6977
Tamhane	培训1天	培训2天	-9.98789 [*]	3.84079	.040	-19.6053	-.3705
		培训3天	-15.69947 [*]	3.17733	.000	-23.8792	-7.5198
	培训2天	培训1天	9.98789 [*]	3.84079	.040	.3705	19.6053
		培训3天	-5.71158	2.56883	.102	-12.2771	.8539
	培训3天	培训1天	15.69947 [*]	3.17733	.000	7.5198	23.8792
		培训2天	5.71158	2.56883	.102	-.8539	12.2771

*. 均值差的显著性水平为 0.05。

表 9-4 为 Tukey B 两两比较输出的结果，它把在 5% 的显著性水平下没有区别的总体放在同一列，作为同类子集。这里，培训 2 天和培训 3 天没有显著区别，它们作为一类。而培训 1 天单独作为 1 类。

表 9-4 同类子集

培训考试得分				
销售培训组		N	alpha = 0.05 的子集	
			1	2
Tukey B ^a	培训1天	20	63.5798	
	培训2天	20		73.5677
	培训3天	20		79.2792

将显示同类子集中的组均值。

a. 将使用调和均值样本大小 = 20.000。

图 9-6 为各个总体的均值的折线图，从中可以直观的看出各个总体均值的趋势。

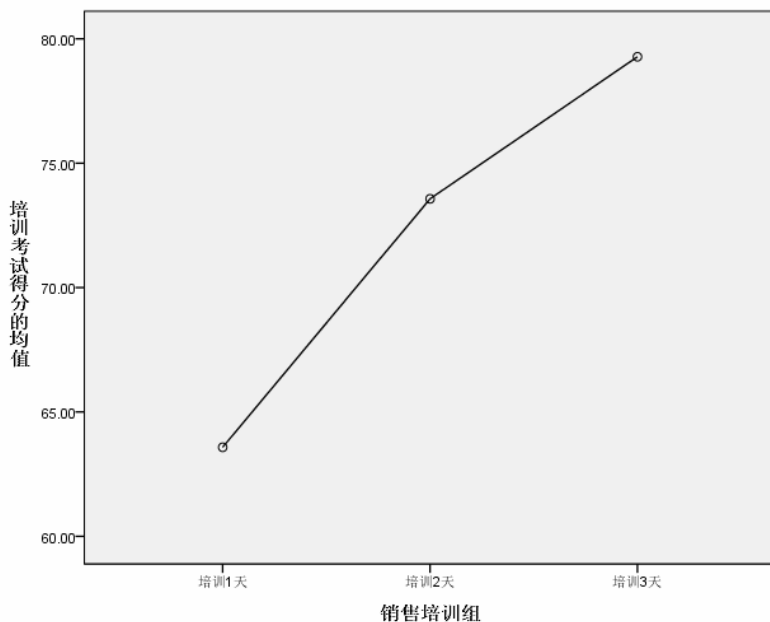


图 9-6 均值图

9.3 多因素方差分析

如果影响试验结果的因素有两个或者两个以上，是否不同的处理对试验结果有显著性影响，不同的因素是否有交互作用？可以应用 SPSS 的一般线性模型（GLM）来完成多因素的方差分析。

9.3.1 多因素方差分析简介

SPSS GLM 过程假设：

- 误差之间相互独立，并且也独立于模型中的其他变量。一般好的试验设计都可以避免违反该条件；

- 不同处理的误差为常数;
- 误差服从均值为 0 的正态分布。

一家连锁零售商店对它们客户的购买习惯进行了一项调查,它记录了客户性别,购买模式、上一个月的购买金额等信息。该商店需要了解在控制客户性别的条件下,是否客户购买的频率和花费的金额有关系,以此来决定是否采取相应的促销活动。

9.3.2 多因素方差分析举例

打开本章的数据文件 `grocery_1month.sav`。

选择【分析】→【一般线性模型】→【单变量】，进入多因素方差分析菜单，如图 9-7 所示。

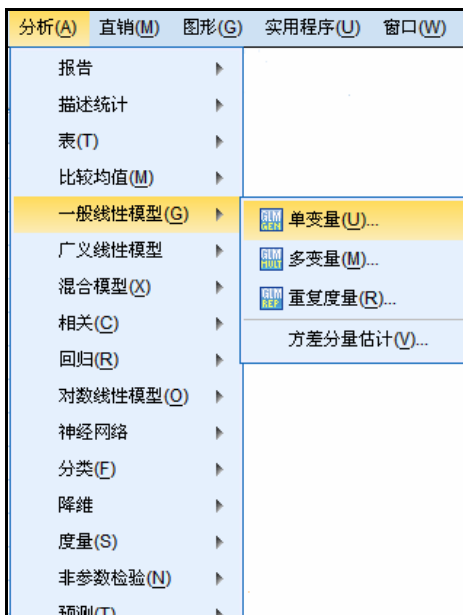


图 9-7 多因素方差分析

在图 9-8 所示的多因素方差分析对话框中,把花费金额变量“`amtspnt`”选入“因变量(D)”框中,把“`gender`”和“`style`”选入“固定因子(F)”框中。单击“绘制(T)”按钮,得到如图 9-9 所示对话框。

在图 9-9 中,把 `style` 选入水平轴, `gender` 选入单图,然后单击“添加”按钮。再把 `style` 和 `gender` 互相交换,选入不同的框中,单击“添加”按钮。完成如图 9-9 所示设置,然后单击“继续”按钮,返回上级对话框图 9-8。在图 9-8 中,单击“保存”按钮,出现如图 9-10 所示的保存对话框,在诊断部分勾选“Cook 距离”和“杠

杆值”。单击“继续”，返回上级对话框。

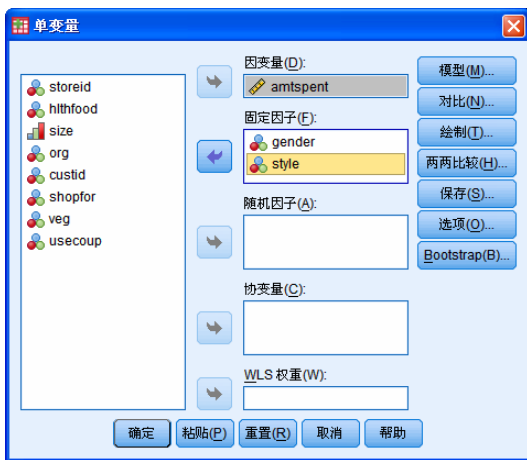


图 9-8 多因素方差分析对话框



图 9-9 绘制对话框

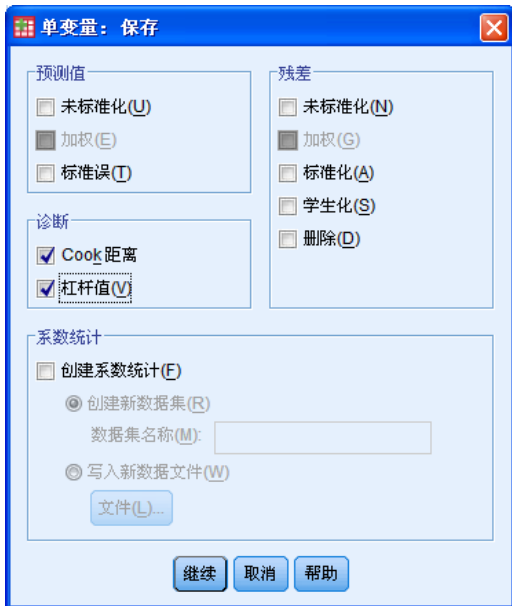


图 9-10 保存对话框

在图 9-8 中，单击“选项”按钮。出现选项对话框，如图 9-11 所示。在输出部分，勾选“描述统计 (D)”、“功效估计”、“方差齐性检验”、“分布-水平图”和“缺乏拟合优度检验 (L)”。单击“继续”，返回上级对话框，如图 9-8 所示。

在图 9-8 中，单击“两两比较”按钮，得到如图 9-12 所示的两两比较对话框。由于 gender 只有两个值，不能进行两两比较，我们把“style”选入两两比较框中。然后勾选“Bonferroni”和“Tukey s-b(K)”两项。单击【继续】按钮，返回如图 9-8 所示的对话框，单击【确定】按钮。



图 9-11 选项



图 9-12 两两比较

以上操作过程可以通过下列的命令语法来实现。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = ' C:\SPSSIntro\Chapter 9\grocery_1month.sav'.
DATASET NAME myData WINDOW=FRONT.
UNIANOVA amtspent BY gender style
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=COOK LEVER
  /POSTHOC=style(BTUKEY BONFERRONI)
```

```

/PLOT=PROFILE(gender*style style*gender)
/PRINT=LOF OPOWER ETASQ HOMOGENEITY DESCRIPTIVE
/PLOT=SPREADLEVEL
/CRITERIA=ALPHA(.05)
/DESIGN=gender style gender*style.

```

在结果查看器中得到表 9-5 到表 9-8 和图 9-13、图 9-14 所示的结果。

表 9-5 的“主体间因子”为方差分析的因子列表，这里有两个因素：性别和购物方式，分别有 2 个水平和 3 个水平，N 为各个因子水平对应的样本中的个案数。

表 9-5 因子列表

主体间因子			
		值标签	N
性别	0	男	185
	1	女	166
购物方式	1	两周一一次；大量	70
	2	每周；类似物品	222
	3	经常；降价物品	59

表 9-6 为描述性统计量，对于男性，每周购物方式消费金额最高，经常购物方式消费金额最低，而对于女性正好相反。从描述性统计量，难以判断不同购物方式消费金额均值之间的差异是由于抽样的随机性所致，还是由于系统性的差别，即不同的购物方式有统计上显著的差别。这需要进一步分析其他的方差分析结果。

表 9-6 描述性统计量

描述性统计量				
因变量:消费额				
性别	购物方式	均值	标准 偏差	N
男	两周一一次；大量	413.0657	90.86574	35
	每周；类似物品	440.9647	98.23860	120
	经常；降价物品	407.7747	69.33334	30
	总计	430.3043	93.47877	185
女	两周一一次；大量	343.9763	100.47207	35
	每周；类似物品	361.7205	90.46076	102
	经常；降价物品	405.7269	80.57058	29
	总计	365.6671	92.64058	166
总计	两周一一次；大量	378.5210	101.25839	70
	每周；类似物品	404.5552	102.48440	222
	经常；降价物品	406.7681	74.42114	59
	总计	399.7352	98.40821	351

表 9-7 中 Levene 检验的 p 值为 0.330，大于 0.05，可以认为满足方差齐性。

表 9-7 方差齐性检验

误差方差等同性的 Levene 检验^a

因变量:消费额

F	df1	df2	Sig.
1.157	5	345	.330

检验零假设, 即在所有组中因变量的误差方差均相等。

a. 设计: 截距 + gender + style + gender * style

表 9-8 为各个因素的效应和交互效应的检验结果。gender 因素的 p 值为 0.000, 小于 0.05, 统计上显著; gender*style 表示 gender 和 style 的交互效应, 它对应的 p 值为 0.017, 所以两个因素的交互效应显著, 即不同的性别在不同的消费方式上所花费金额的模式是不同的。

表 9-8 模型主体间效应的检验

主体间效应的检验

因变量:消费额

源	III 型平方和	df	均方	F	Sig.
校正模型	469402.996 ^a	5	93880.599	11.092	.000
截距	3.936E7	1	3.936E7	4650.274	.000
gender	158037.442	1	158037.442	18.672	.000
style	33506.210	2	16753.105	1.979	.140
gender * style	69858.325	2	34929.163	4.127	.017
误差	2920058.824	345	8463.939		
总计	5.948E7	351			
校正的总计	3389461.820	350			

a. R 方 = .138 (调整 R 方 = .126)

图 9-13 和图 9-14 为两种因素不同水平组合的均值图。

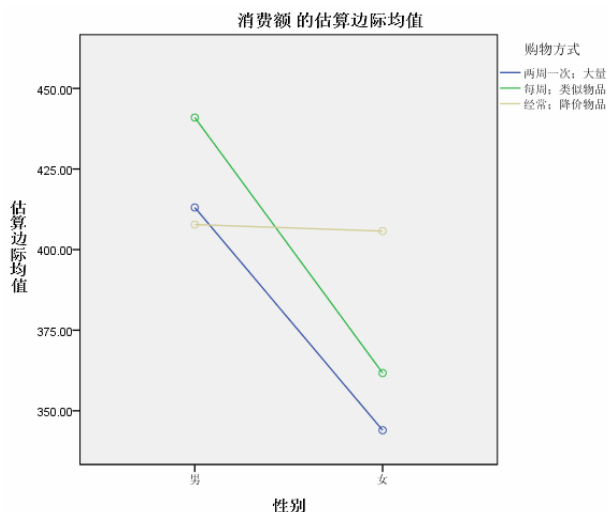


图 9-13 gender*style 均值图

男性和女性在每周购物和两周一次购物的均值线是平行的，都是男高女低；而在经常购物上，二者差距不大，经常购物均值线和另外两条线有交叉，表明二个因素有交互效应。效应是否显著在“主体间效应的检验”表中标识。

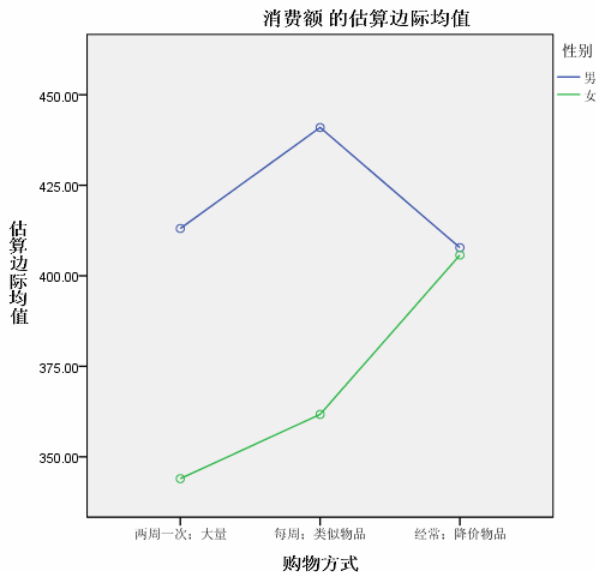


图 9-14 style*gender 均值图

从 style*gender 均值图看出，男性的所有消费方式的消费金额均大于女性，男性和女性消费方式的曲线是不平行的，表明二者有交互效应。女性在经常性购物中花费金额最多；而男性则在每周购物方式中花费最多。

综上所述，性别因素在消费金额中有显著的效应，而消费方式因素则不起显著作用；这两个因素具有交互效应，一周购买一次的男性顾客比经常购买的顾客更能给商店带来利润，女性顾客恰好相反。商店可以据此制定相应的营销策略。

注意：SPSS 的多因素方差分析中允许含有随机因素。除了控制因素以外，有时候一些和研究的问题没有直接关系的因素也会对试验结果有影响，这种因素被称为随机因素。在试验设计时，尽可能地考虑到对实验结果有影响的因素，并收集相关数据，参与到分析中，可能会提高模型的解释性。

9.4 协方差分析 (ANCOVA)

9.4.1 协方差分析简介

在方差分析中，无论是单因素方差分析还是多因素方差分析，控制因素是可以

控制的,其水平可以在试验中人工控制和确定,并且控制因素的所有水平会全部在试验中出现。但是,实际问题中,有些控制因素很难人为控制,比如比较失业人群经过不同的再培训项目后的工资水平是否有差异,失业人群在失业以前的工资水平是无法控制的,但是它会对培训后的工资水平有影响。又比如,比较三种不同饲料对肉鸡体重的影响,在用这些饲料喂养前,肉鸡的体重就有差异,它们肯定会影响到用饲料饲养后肉鸡的体重。影响试验结果但是又无法人工确定和控制其水平的因素被称为协变量(Covariate),协变量一般为尺度类型数据。既分析控制因素影响,又分析协变量的影响以及控制因素和协变量关系的方法称为协方差分析。

协方差分析是针对在试验阶段难以控制或者无法严格控制的因素,在统计分析阶段进行统计控制,它在扣除协变量的影响后再对修正后的主效应进行方差分析,是一种把直线回归和方差分析结合起来的方法。协方差分析的数学模型为:

$$y_{ij} = \mu + a_i + \beta z_{ij} + \varepsilon_{ij}, i = 1, 2, \dots, k; j = 1, 2, \dots, n.$$

这里 y_{ij} 表示在控制因素的水平 i 下的第 j 次试验的因变量观测值; μ 为因变量总体均值; a_i 表示控制因素的水平 i 下对因变量产生的效应; β 为协变量的回归系数; z_{ij} 表示在控制因素的水平 i 下的第 j 次试验的协变量观测值; ε_{ij} 为抽样误差,假设它是服从方差相等的正态分布变量。

9.4.2 协方差分析案例分析

政府就业促进部门想了解他们的就业促进项目是否发挥了实质性的作用,他们随机选取了参加该项目的人和没有参加该项目的人,调查这些人在实施该项目前后的收入变化,调查结果保存于数据文件 `workprog.sav` 中。这里研究的目标变量为参加项目后人们的薪水,用它来衡量人们找到工作的好坏,即变量“`incaft`”(参加项目后的薪水)为因变量。由于参加该就业促进项目之前人们的薪水是不同的,如果不考虑该因素,直接比较参加项目之后人们薪水的区别是不合理的。因此把参加该项目前人们的薪水(`incbef`)作为协变量,把是否参加就业促进项目(`prog`)作为控制因素,即自变量。协方差分析(ANalysis Of COVariance, ANCOVA)除了方差分析的假设条件之外,协方差分析还要求在控制因素和协变量之间没有显著的交互作用。

1. 协变量和因变量、协变量和控制因素的关系

在进行协方差分析之前,一般要先检查进行协方差分析的前提条件是否满足如

下情况：

- 协变量和因变量之间是否有线性关系；
- 可以通过绘制散点图来直观的观测二者之间线性关系的强弱；
- 控制因素和协变量之间是否有交互作用。

可以先预先进行方差分析检查二者之间的交互效应是否显著。

打开本章的数据文件 `workprog.sav`，首先绘制协变量—参加项目前的薪水和因变量—参加项目后的薪水之间的散点图，这里用不同的图案区别不同控制因素水平下的散点图，如图 9-15 所示。

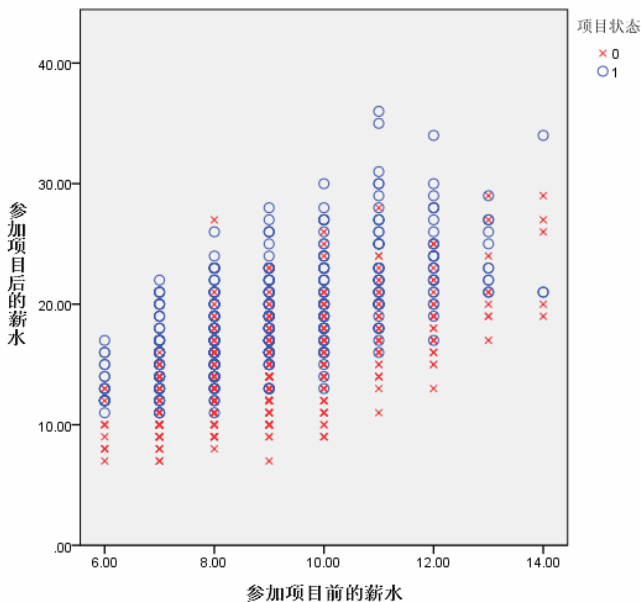


图 9-15 协变量和因变量的散点图

从散点图可知，在参加就业项目和没有参加就业项目状态下，协变量和因变量分别具有线性关系，随着参加项目前的薪水的增大，参加项目后的薪水也增大。

下面检查控制因素和协变量之间是否具有交互作用。

选择【分析】→【一般线性模型】→【单变量】，出现如图 9-16 所示的单变量对话框。把“`incaft`”选入“因变量(D)”框中；把变量“`prog`”选入“固定因子(F)”框中，把“`incbef`”选入“协变量(C)”框中。



图 9-16 协方差分析

单击【模型】按钮，得到如图 9-17 所示的模型选择对话框。在“构建项”中选择“交互”，在对话框左侧的“因子与协变量”框中，同时选中“incbef”和“prog”，单击向右的箭头，把它们选入到右侧的“模型”框中。然后在“构建项”中选择“主效应”，在对话框左侧的“因子与协变量”框中，分别把“incbef”和“prog”选入到右侧的“模型”框中，设置完成后如图 9-17 所示。

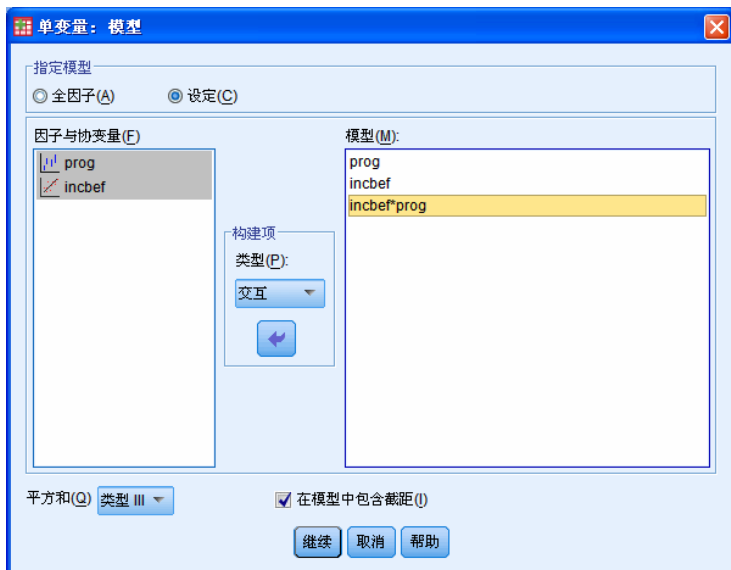


图 9-17 指定模型

单击【继续】按钮，返回上级对话框图 9-16，然后单击“选项”按钮，得到如图 9-18 所示的选项对话框，在“输出”部分勾选“功效估计 (E)”。

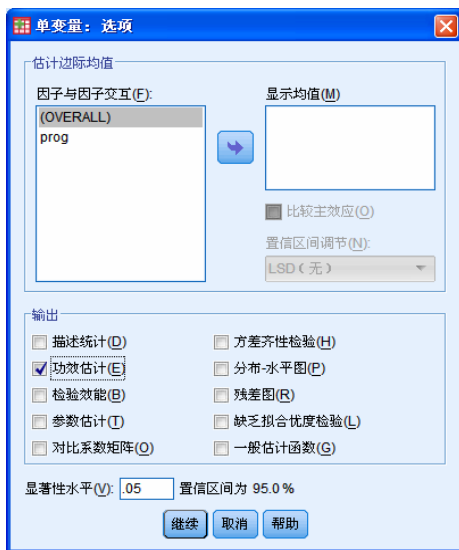


图 9-18 选项对话框

单击【继续】按钮，返回上级对话框，如图 9-16 所示，然后单击【确定】按钮，即完成协方差分析的设置。上述操作可以通过下列语法命令来实现。

```
NEW FILE.
DATASET CLOSE ALL.
GET FILE = ' C:\SPSSIntro\Chapter 9\workprog.sav'.
DATASET NAME myData WINDOW=FRONT.
UNIANOVA incbef BY prog WITH incbef
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=ETASQ
  /CRITERIA=ALPHA(.05)
  /DESIGN=prog incbef incbef*prog.
```

在结果查看器中得到如表 9-9 所示的主体间效应检验结果。

表 9-9 主体间效应的检验

主体间效应的检验

因变量:参加项目后的薪水

源	III 型平方和	df	均方	F	Sig.	偏 Eta 方
校正模型	12295.033 ^a	3	4098.344	429.755	.000	.564
截距	131.271	1	131.271	13.765	.000	.014
prog	106.795	1	106.795	11.199	.001	.011
incbef	7152.586	1	7152.586	750.025	.000	.430
prog * incbef	4.292	1	4.292	.450	.502	.000
误差	9498.318	996	9.536			
总计	297121.000	1000				
校正的总计	21793.351	999				

a. R 方 = .564 (调整 R 方 = .563)

从表 9-9 知,控制因素和协变量的交互项“ $\text{prog}*\text{incbef}$ ”对应的显著性值为 0.502,大于 0.05,并且其偏 Eta 方为 0,说明交互项对因变量的影响可以忽略不计。所以,控制因素和协变量之间的交互效应统计上不显著。基于此,可以认为参加项目前的收入(协变量)与是否参加该项目(控制因素)无关。

综上所述,协变量和控制因素满足进行协方差分析的条件。

2. 协方差分析

打开本章的数据文件 `workprog.sav`,选择【分析】→【一般线性模型】→【单变量】。把“`incaft`”选入“因变量(D)”框中;把变量“`prog`”选入“固定因子(F)”框中,把“`incbef`”选入“协变量(C)”框中。设置如图 9-19 所示。



图 9-19 协方差分析

单击【模型(M)】按钮,得到如图 9-20 所示的设置因子模型对话框。可以选择和试验所对应的因子模型,这里保留默认值“全因子模型”。



图 9-20 设置因子模型

单击【继续】按钮，返回上级对话框如图 9-19 所示。单击“选项”按钮，如图 9-21 所示。



图 9-21 设置选—选择输出和显示调整后边际均值

在输出部分，勾选“描述性统计 (D)”、“功效估计 (E)”、“参数估计 (T)”、“方差齐性检验 (H)”和“分布-水平图 (P)”。选中“因子与因子交互 (F)”框中的“prog”变量，单击中间向右箭头，把“prog”选入右侧的“显示均值 (M)”框中。勾选“比较主效应 (O)”选项，在“置信区间调节 (N)”下方的列表框中

选择“Bonferroni(B)”（也可以选择另外两个置信区间调节选项—SIDAK 和 LSD）。如图 9-21 所示。

单击【继续】按钮，返回上级菜单，如图 9-19 所示，然后单击【确定】按钮，完成协方差分析的设置。以上操作过程可以通过下列语法命令来完成。

```
UNIANOVA incaft BY prog WITH incbef  
  /METHOD=SSTYPE(3)  
  /INTERCEPT=INCLUDE  
  /EMMEANS=TABLES(prog) WITH(incbef=MEAN) COMPARE ADJ(BONFERRONI)  
  /PRINT=PARAMETER ETASQ HOMOGENEITY DESCRIPTIVE  
  /PLOT=SPREADLEVEL  
  /CRITERIA=ALPHA(.05)  
  /DESIGN=incbef prog.
```

在结果查看器中，得到结果如表 9-10 到表 9-15 所示。

表 9-10 描述性统计量和方差齐性检验

描述性统计量			
因变量:参加项目后的薪水			
项目状态	均值	标准 偏差	N
0	14.4023	3.89303	517
1	18.9379	4.28162	483
总计	16.5930	4.67067	1000

误差方差等同性的 Levene 检验 ^a			
因变量:参加项目后的薪水			
F	df1	df2	Sig.
4.873	1	998	.028
检验零假设，即在所有组中因变量的误差方差均相等。			
a. 设计: 截距 + incbef + prog			

从表 9-10 的“描述性统计量”部分可知，参加项目前和参加项目后的均值是不同的。从“误差方差等同性的 Levene 检验”部分知，Levene 检验的显著性值为 0.028，小于 0.05。由于因素的水平数只有两个，从描述性统计量看出这两个水平的标准偏差差距不是太大，所以不能够武断地得出方差不等的结论。结合图 9-22 的水平-分布图，看出分布的跨度小于 0.4（即 4.3~3.9），而水平的跨度大于 4.5（即 19~14.5）。即薪水的变差相对于其均值较小，假设方差齐性是比较安全的。

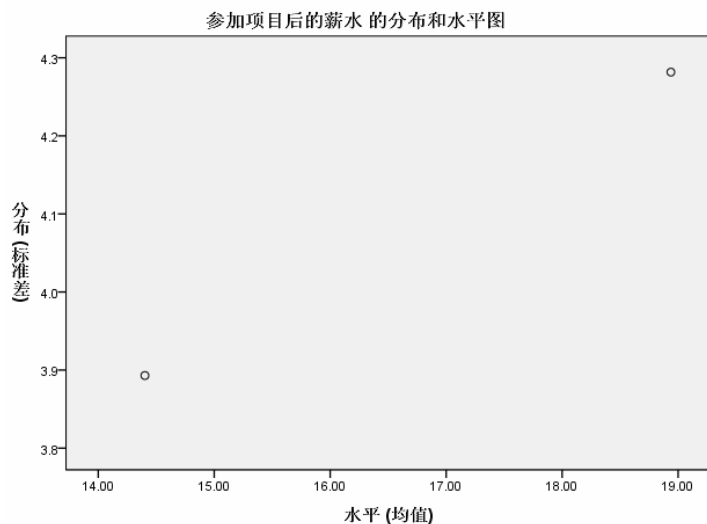


图 9-22 水平-分布图

表 9-11 主体间效应的检验

因变量: 参加项目后的薪水

源	III型平方和	df	均方	F	Sig.	偏 Eta 方
校正模型	12290.741 ^a	2	6145.370	644.763	.000	.564
截距	131.400	1	131.400	13.786	.000	.014
incbef	7153.844	1	7153.844	750.571	.000	.429
prog	4735.662	1	4735.662	496.859	.000	.333
误差	9502.610	997	9.531			
总计	297121.000	1000				
校正的总计	21793.351	999				

a. R 方 = .564 (调整 R 方 = .563)

从表 9-11 主体间效应检验可知, 协变量和控制因素都对因变量有显著的影响。这两个因素对因变量(参加项目后的薪水)效应的大小可以从表 9-12 的参数估计值得到。

注意到表 9-12 控制因素 prog 前的系数为-4.357, 意味着没有参加项目的人的薪水比参加项目的人的少\$4,357。

表 9-12 参数估计

因变量: 参加项目后的薪水

参数	B	标准误差	t	Sig.	95% 置信区间		偏 Eta 方
					下限	上限	
截距	4.197	.556	7.548	.000	3.106	5.288	.054
incbef	1.636	.060	27.397	.000	1.519	1.753	.429
[prog=0]	-4.357	.195	-22.290	.000	-4.741	-3.974	.333
[prog=1]	0 ^a

a. 此参数为冗余参数, 将被设为零。

表 9-13 为把参加项目和没有参加项目两个总体进行 Bonferroni 调整, 把协变量

一参加项目前的薪水调整到同一个水平 8.9540 进行参数估计。这里，调整协变量后参加项目和没有参加项目的均值估计值分别为 14.488 和 18.846。注意和表 9-10 没有经过协变量调整的估计值（分别为 14.4023 和 18.9379）进行比较。

表 9-13 调整协变量后的均值估计

因变量:参加项目后的薪水

项目状态	均值	标准误差	95% 置信区间	
			下限	上限
参加项目	14.488 ^a	.136	14.222	14.755
没有参加项目	18.846 ^a	.141	18.570	19.121

a. 模型中出现的协变量在下列值处进行评估: 参加项目前的薪水 = 8.9540.

表 9-14 是经过 Bonferroni 调整后的配对比较，这里控制因素只有两个水平，它们的差值的显著性值为 0.000，小于 0.05，因此参加项目和没有参加项目两个组别的参加项目后薪水是有显著区别的。

表 9-14 Bonferroni 成对比较

因变量:参加项目后的薪水

(I) 项目状态	(J) 项目状态	均值差值 (I-J)	标准误差	Sig. ^a	差分的 95% 置信区间 ^a	
					下限	上限
参加项目	没有参加项目	-4.357 [*]	.195	.000	-4.741	-3.974
没有参加项目	参加项目	4.357 [*]	.195	.000	3.974	4.741

基于估算边际均值

*. 均值差值在 .05 级别上较显著。

a. 对多个比较的调整: Bonferroni。

表 9-15 是对参加项目后的薪水的单变量检验。

表 9-15 单变量检验

因变量:参加项目后的薪水

	平方和	df	均方	F	Sig.	偏 Eta 方
对比	4735.662	1	4735.662	496.859	.000	.333
误差	9502.610	997	9.531			

F 检验 项目状态 的效应。该检验基于估算边际均值间的线性独立成对比较。

9.5 小结

方差分析本质上是多个总体均值的比较。根据控制因素个数的不同，方差分析分为单因素方差分析、多因素方差分析等。本章介绍了方差分析的基本思想、术语，单因素方差分析、多因素方差分析的方法和技巧。另外，本章同时介绍了协方差分析的方法和技巧。

思考与练习

1. 一家关于 MBA 报考、学习、就业指导的网站希望了解国内 MBA 毕业生的起薪是否与各自所学的专业有关,为此,他们在已经在国内商学院毕业并且获得学位的 MBA 学生中按照专业分别随机抽取了 10 人,调查了他们的起薪情况,数据文件为 MbaSalary.sav,根据这些调查他们能否得出专业对 MBA 起薪有影响的结论?
2. 美国得克萨斯州的一所大学提出了三种 GMAT 辅导课程:即 3 小时复习、1 天课程和 10 周强化班,他们需要了解这三种辅导方式如何影响 GMAT 成绩。另外,通常考生来自三类院校,即商学院、工学院、艺术与科学院。因此,了解不同类型学校毕业的考生 GMAT 成绩是否有差异也是一个让人感兴趣的话题。他们在三类学校中每一个随机抽取 6 个学生,随机指派两名到一门辅导课程中,最后他们的 GMAT 成绩结果记录于数据文件 GmatScore.sav 中。问题为:
 - 1) 不同的辅导课程是否对学生 GMAT 的成绩有显著的影响?来自不同类型学校的学生的 GMAT 成绩是否有显著的差别?请给出理由。
 - 2) 是否一类学校的考生适应一种辅导课程,而另一类学校的考生适合其他课程?请给出理由。
3. 为研究三种不同饲料 A1、A2 和 A3 对生猪体重增加(wyh)的影响,将生猪随机分成三组喂养不同的饲料(sl)。由于生猪体重的增加理论上会受到喂养前的体重影响,而喂养前的体重则是难以控制的,相关的试验数据记录在文件 Anocov.sav 中。请用协方差分析的方法比较三种鸡饲料在增加生猪体重上是否有显著差别。

参考文献

1. 梁冯珍,关静等译.统计学(第5版).北京:机械工业出版社,2009。
2. 薛薇.SPSS 统计分析方法及应用(第二版).北京:电子工业出版社,2008。

SPSS 输出管理器简介

本章学习目标：

- 输出管理器显示和隐藏结果；
- 移动、复制和删除结果、添加和编辑文本；
- 设置缺省选项；
- 应用枢轴表编辑器；
- 打印输出管理器的内容；
- 导出 SPSS 结果到其他应用程序。

SPSS 统计分析的结果和 SPSS 绘图结果会在 SPSS 输出查看器中显示（SPSS Statistics Viewer）。SPSS 结果浏览器中的内容可以是统计表格、统计图形和文本，取决于运行的统计程序和相应的选项设置。在 SPSS 结果浏览器可以轻松地浏览特定的统计结果，可以管理和操作选定的输出内容。也可以生成包含所需要的结果的 Word 文档、PPT 报告或者 Excel 电子表格。本章中将介绍如何在 SPSS 结果浏览器中找到所需要的结果，如何组织 SPSS 的输出，如何对已经生成的结果进行编辑以适应报告的需要。

10.1 SPSS 结果浏览器简介

SPSS 结果浏览器分为两个区域：目录区域（或称为大纲区域）和内容区域，如图 10-1 所示。结果浏览器的左边部分为目录区域，它就如一本书的目录，从这里可以一目了然整个 SPSS 的输出结果，可以方便地定位到用户需要的输出内容。结果浏览器的右边部分为内容区域，它是用户运行 SPSS 程序所输出的详细内容，包括统计图形、统计表格、分析结果的标题、结果注释、运行 SPSS 程序产生的警告或者错误信息，如果进行了相关设置，结果浏览器还可以输出所运行的程序的语法命令。

SPSS 结果浏览器中的内容可以保存为后缀为 .SPV 的文件，供以后应用。从 SPSS

版本 16 开始, SPSS 结果文件的后缀名为.SPV。而 SPSS 版本 15 或者以前版本的结果文件的后缀为.SPO, 它们不能在 SPSS 版本 16 或者以后的版本中打开。如果需要阅读 SPSS 版本 15 或者以前版本生成的结果文件, 需要相应的 SPSS 版本或者安装 SPSS 公司免费提供的结果浏览器软件 Legacy Viewer (又称为 Smart Viewer, 在本书的技术支持网站 books.minewin.com 可以下载该浏览器)。

注意: SPSS 版本 16 或者以后版本的结果文件不能在 SPSS 版本 16 以前的 SPSS 统计软件阅读, 克服的办法是把 SPSS 版本 16 或者以后版本生成的结果导出为 Word 或者 PDF 文档, 然后借助第三方的软件来阅读这些结果。

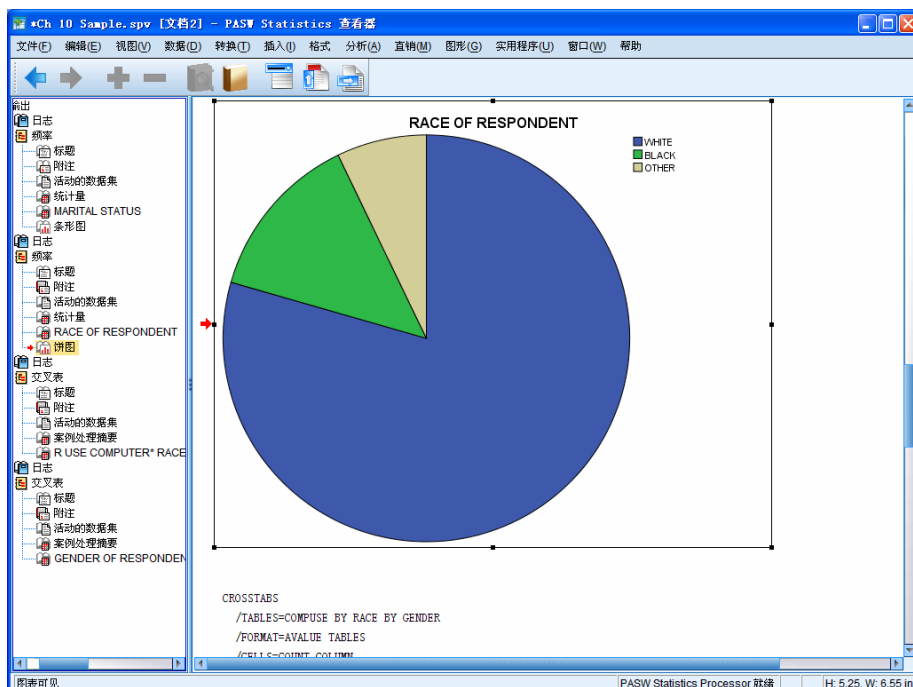


图 10-1 SPSS 结果浏览器

10.2 浏览和编辑 SPSS 的分析结果

SPSS 结果浏览器的内容是一系列按照层次结构组织的对象组成的, 每个表格或者统计图都是一个对象, 另外还有一些单独的对象, 例如输出内容的标题、日志等。每一个统计程序的输出都是许多对象构成的, 称为对象块。SPSS 的目录区域显示每个单独的对象、对象块和其下属的对象。

10.2.1 目录区域的对象

SPSS 浏览器的目录区域部分的每个对象块都有一个目录形状的图标 () 或者

和标题。如图 10-2 所示。SPSS 结果浏览器的目录区域可以看作一本书的目录，最上层的目录为“输出”标题，它表示这是一次或者几次 SPSS 会话的输出；它的下一层即为二级目录，分别对应每次运行 SPSS 程序所产生的结果，它的下一级为三级目录，为该 SPSS 程序所产生的所有结果。

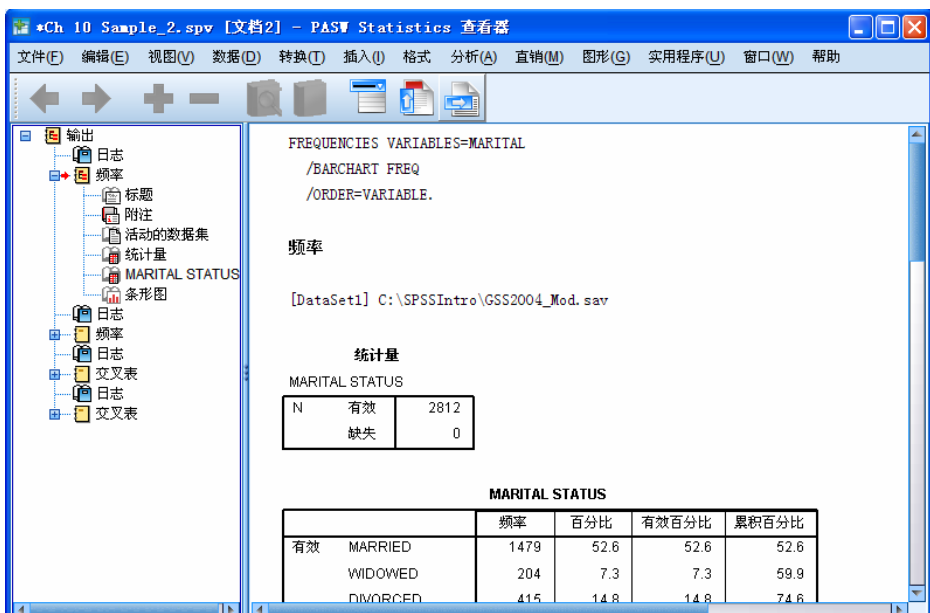







图 10-2 SPSS 输出浏览器的目录区域

如果把“输出”作为一棵树的根的话，每个程序所产生的结果组成的对象块即为此棵树的子分支，对象块的对象为树的叶子。

：该部分是运行某个 SPSS 过程产生的所有结果所组成的对象块，类似一本书的二级目录。该图标的左边是一个“—”，表示该对象块已经全部展开了。如果单击该“—”，则该对象块的下一级目录将被隐藏，图标变为，图标左边的“—”将变为“+”。

：和上边的图标是一样的，它代表运行某个 SPSS 过程产生的所有结果所组成的对象块，类似一本书的二级目录。该图标的左边是一个“+”，表示该对象块还没有展开，单击该“+”则展开该对象块，显示它的下一级目录，展示该对象块下的所有内容，同时图标变为。

：它标识运行每个 SPSS 程序的日志，即语法命令、警告和出错信息等。


在每个对象块下有许多对象，它们是所运行的 SPSS 程序的输出，所有对象块都包括如下的标题、附注、活动的数据集这三个对象：


- 标题是该程序产生的输出的总标题；
- 附注是对所运行的程序的注释，包括该程序的输入数据、当前活动的数据集、是否应用了过滤器、是否设置了权重、是否有文件拆分、对缺失值的处理方法、该程序的语法命令、该程序所花费的计算机时间等，如图 10-3 所示；
- 活动的数据集：标识该程序所分析的数据集，或者说当前输出所对应的数据集。


附注		
创建的输出		15-四月-2010 17时46分56秒
注释		
输入	数据	C:\SPSSIntro\GSS2004_Mod.sav
	活动的数据集	DataSet1
	过滤器	<none>
	权重	<none>
	拆分文件	<none>
	工作数据文件中的 N 行	2812
缺失值处理	对缺失的定义	用户定义的缺失值被视为缺失。
	使用的案例	每个表的统计量基于各表内对指定范围内所有变量都具有有效数据的所有案例。
语法		CROSSTABS /TABLES=GENDER BY DEGREE /FORMAT=AVALUE TABLES /CELLS=COUNT ROW /COUNT ROUND CELL.
资源	处理器时间	00:00:00.016
	已用时间	00:00:00.375
	要求的维数	2
	可用单元格	174762


图 10-3 附注

双击每个对象块下的对象所对应的图标，可以显示该对象或者隐藏该对象。每个对象前的图标都是一本书，附上一个标识该对象内容的小图标，分别用于辨识该对象为文本对象、数值对象或者图形对象。如果对象前的图标是敞开状态，标识该对象是显示状态；否则，该对象是隐藏状态。对象前的图标有两类，每类有两种状态：显示或隐藏，共计下列 4 种图标：

：为表格对象，该对象为显示状态。双击后，该表格将在内容区域被隐藏；

：为表格对象，该对象为隐藏状态。双击后，该表格将在内容区域被显示；

：为图形对象，该对象为显示状态。双击后，该图形将在内容区域被隐藏；

：为图形对象，该对象为隐藏状态。双击后，该图形将在内容区域被显示。

除了以上三类对象以外，每个 SPSS 输出都有它们自己的“特殊对象”。例如，“频率”过程有统计量、频数表和条形图等对象；“交叉表”过程有活动的数据集、案例处理摘要、交叉表等对象。

注意：

- 当单击目录区域的某个图标时，如果该图标的内容没有被隐藏，将在内容区域定位到该图标所对应的输出内容。这时，在目录区域的图标前有一个红色小箭头，相应的内容区域部分将被选中，且其右端有一个红色箭头。如图 10-4 所示。
- 如果图标的内容被隐藏，内容区域没有任何显示。
- 如果结果浏览器中有大量结果，隐藏部分结果而不是实际删除它们，这可以在保留所有结果的基础上方便的关注感兴趣的结果。同时，方便选择需要打印的输出（打印“所有可见”内容）或者导出“所有可见”内容。

运行同一个过程多次，在结果浏览器中会相应的产生多次结果输出，并且它们的对象名称（或者标题）都完全一样。这样就很难分清哪个输出结果对应那一次程序运行，不同运行的设置有何不同。经过一段时间后，用户很难找到需要的结果。因此，在保存输出结果之前，需要修改输出结果的标题或者添加相关的注释文字，或者重新组织输出结果的结构。一个好的习惯是每次 SPSS 给出输出结果后，如果该结果需要保存以供后用，先把对象块的名称以及其下属的多个对象的名称进行编辑，变为自己容易记住的名称。如图 10-4 所示，单击“RACE OF RESPONDENT”部分，在目录区域和内容区域相应的内容被选中，为了便于记忆，把“RACE OF RESPONDENT”重新命名为“响应者的性别频率表”。

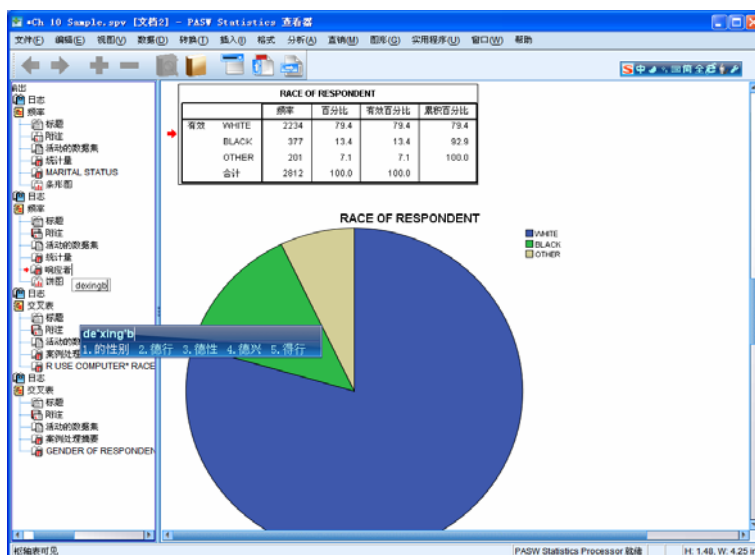


图 10-4 修改对象的标题

枢轴表 商店ID* 商店组织 交叉制表

文件(F) 编辑(E) 视图(V) 插入(I) 透视(P) 格式(O) 帮助(H)

商店ID* 商店组织 交叉制表

统计量 计数

	商店组织				合计
	重点是生产	重点是熟食	重点是面包	没有重点	
36	0	0	0	8	8
37	0	12	0	0	12
38	0	0	0	5	5
39	0	0	0	10	10
40	0	5	0	0	5
41	0	0	0	5	5
42	0	0	0	7	7
43	6	0	0	0	6
44	0	0	0	5	5
45	0	0	0	3	3
46	4	0	0	0	4
47	6	0	0	0	6
48	4	0	0	0	4
49	0	0	5	0	5
50	0	0	0	6	6
51	0	0	6	0	6
52	3	0	0	0	3
53	0	6	0	0	6

图 10-6 枢轴表编辑器

10.2.3 移动、复制和删除结果

通过目录区域可以调整输出结果的显示顺序、删除结果和复制结果。

1. 移动结果

在目录区域选中需要移动的对象，用鼠标拖动到需要的位置，然后释放即可。

2. 删除结果

在目录区域选中需要删除的对象，然后按【Delete】键即可。或者，选中需要删除的对象，单击鼠标右键，选择【剪切】，也可以删除对象。

3. 复制结果

在目录区域选中需要复制的对象，选择【编辑(E)】→【复制】，然后移动光标到需要复制到的位置，选择【编辑(E)】→【之后粘贴】，则复制的内容将出现在相应位置之后。

10.2.4 添加和编辑文本

在 SPSS 输出中，用户可以添加注释、描述性文字。也可以添加新标题、添加分页符等。

选择【插入】→【分页符(P)】，将在输出中添加分页符，打印或者打印预览时，将在分页符位置进行分页。

选择【插入】→【标题】，将在输出管理器的目录区域建立一个标题，用于组织管理目录区域的内容。

选择【插入】→【新建文本】，可以添加对结果的描述性内容。

选择【插入】→【新建页面标题】，将对新的页面添加标题。

在【编辑】→【选项】菜单中的“查看器”标签中，可以设定输出浏览器中的日志、警告、附注、标题、页面标题等文字的显示方式，例如字体、尺寸（即字号）、颜色等。用户可以根据自己的喜好进行客户化设置，如图 10-7 所示。

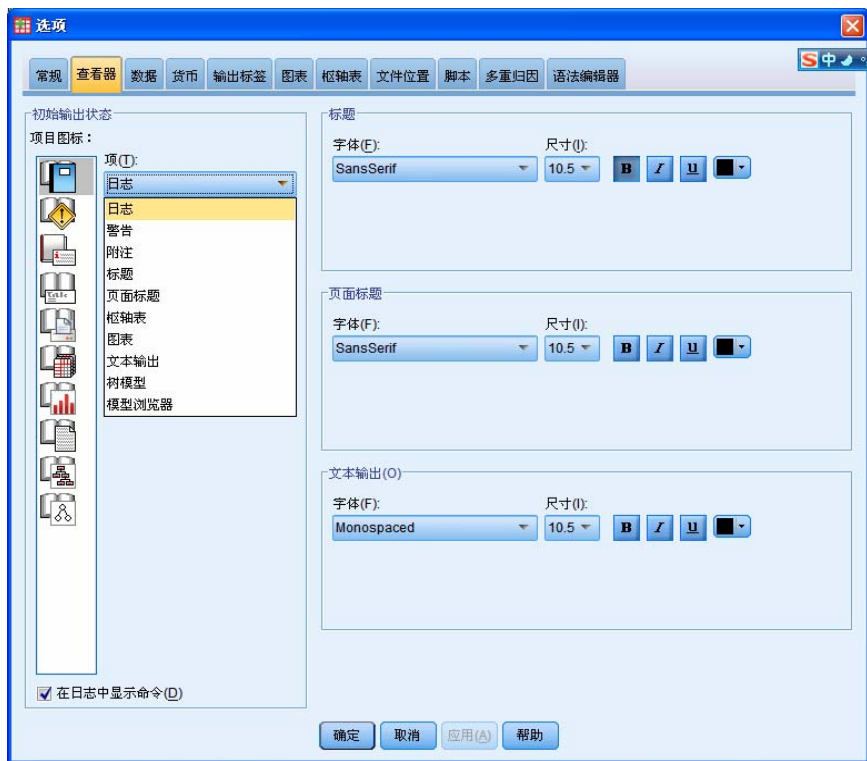


图 10-7 设置输出管理器中文本的显示方式

在 SPSS【选项】菜单的“查看器”标签中，“初始输出状态”部分设置十个输出对象在 SPSS 输出查看器中初始状态为显示或隐藏。如果设置为隐藏，则只有对象标题在目录区域中显示，而内容区域不显示。如图 10-8 所示。

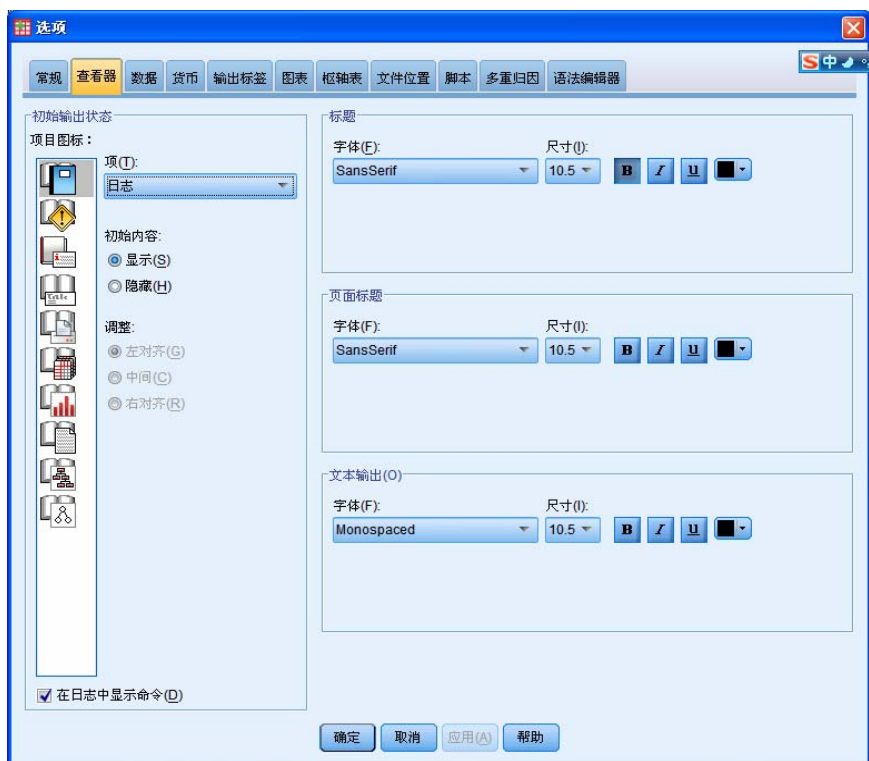


图 10-8 设置 10 个输出项目

这 10 个对象分别为：

- 日志：内置的 SPSS 命令语法；
- 警告：与 SPSS 操作和数据相关的警告信息；
- 附注：SPSS 运行过程的详细信息，默认隐藏；
- 标题：输出项目的标题；
- 页面标题：新插入页面的标题；
- 枢轴表：输出的枢轴表对象；
- 图表：所有输出的图表；
- 文本输出：文本格式输出的对象或者插入的文本；
- 树模型：SPSS 决策树模块产生的输出结果；
- 模型浏览器：SPSS17 或者以后版本中最近邻聚类所产生的输出结果。

10.3 枢轴表编辑器

和 Excel 或者其他 BI 软件类似，SPSS 可以产生枢轴表，并能对枢轴表进行编辑。原则上，SPSS 所有输出都可以在枢轴表中显示数值。枢轴表中的对象或者元素

可以由枢轴表编辑器来编辑和重新排列。在枢轴表编辑器中，可以对表格进行转置，也可以改变表格的边框、格式等属性。可以把客户化后枢轴表的枢轴表格式信息另存为 TableLook 文件，然后在其他表格中应用该文件，这样可以把该种表格格式应用所有表格中。

选中需要编辑的表格，单击鼠标右键，弹出一个弹出窗口。然后选择【编辑内容】，其下有两个选项：“在阅读器中”或者“在单独窗口中”，前者不产生新的窗口，而后者则单独在一个枢轴表编辑器中编辑。如图 10-9 所示。

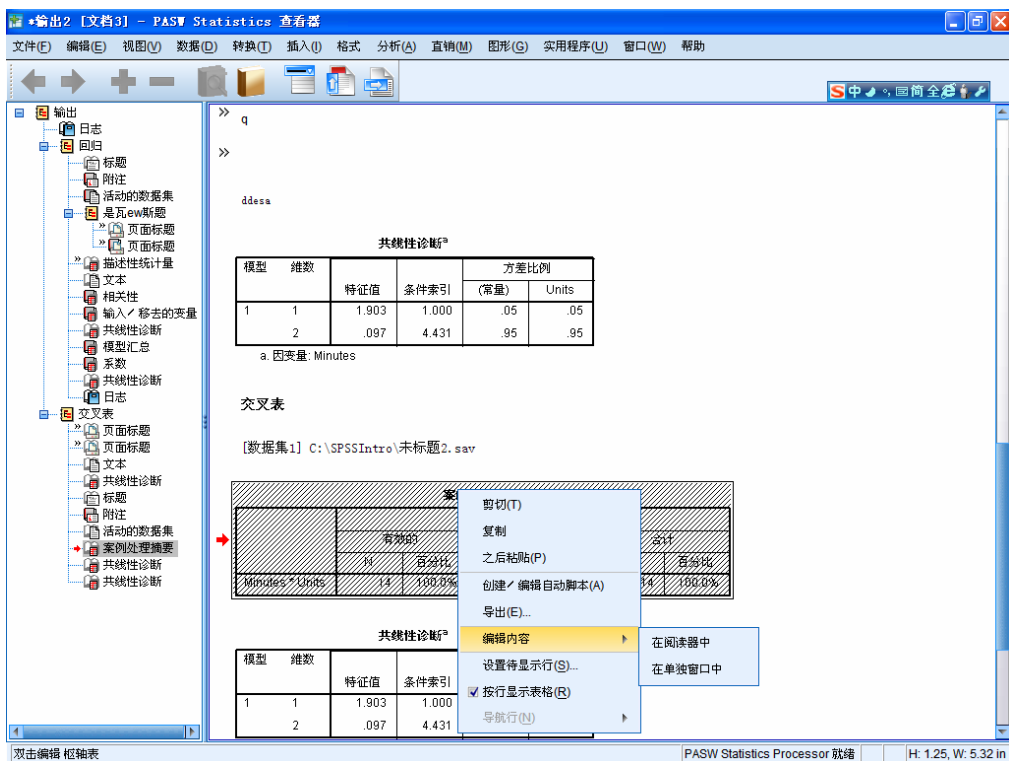


图 10-9 编辑对象

打开本章的示例输出文件 Ch 10 Sample.spv，选择“R USE COMPUTER*RACE OF RESPONDENT*GENDER OF RESPONDENT 交叉制表”对象，如图 10-10 所示。

单击鼠标右键，选择【编辑内容】→【在单独窗口中】，如图 10-11 和图 10-12 所示。

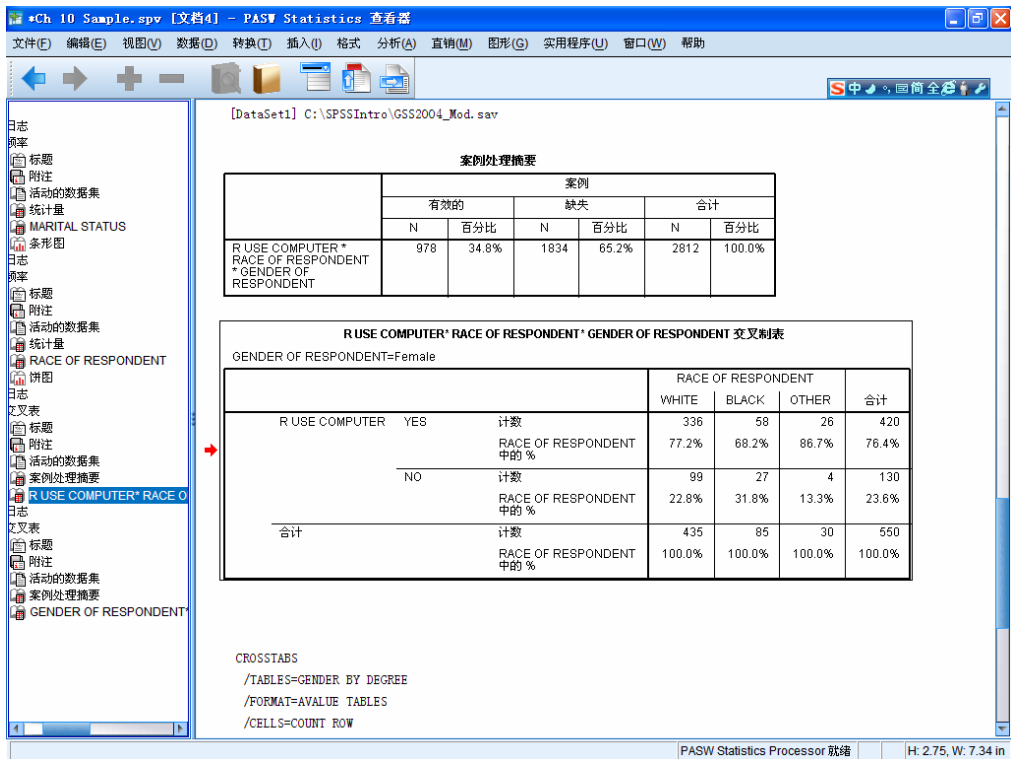


图 10-10 枢轴表

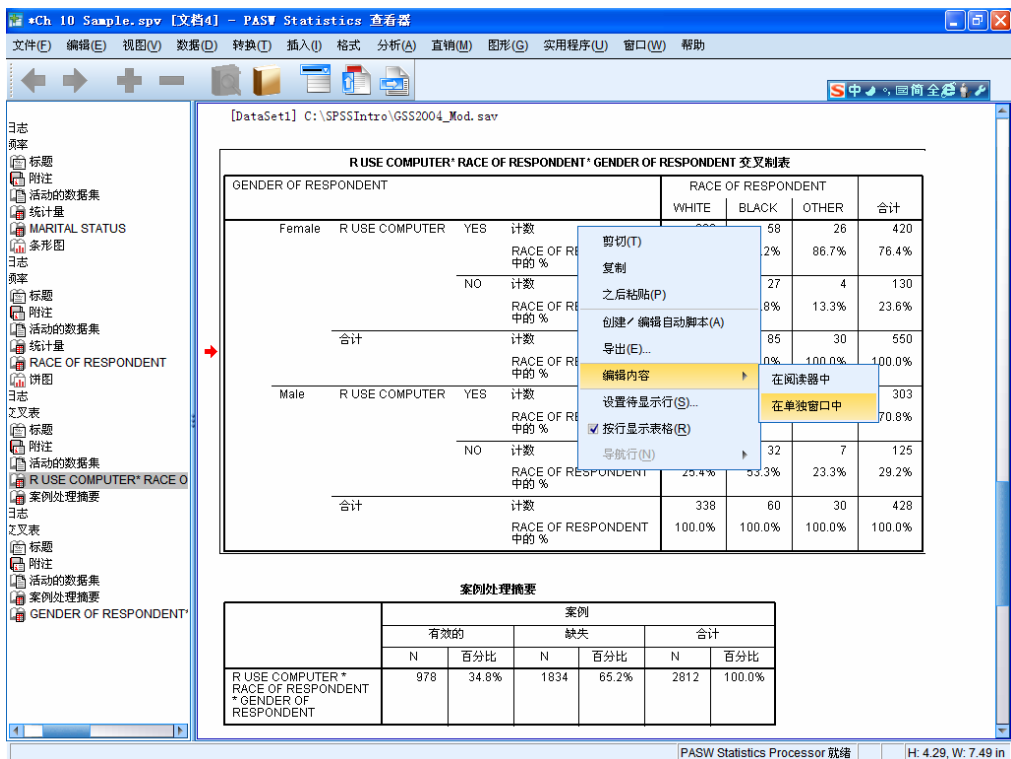


图 10-11 选择枢轴表

枢轴表

R USE COMPUTER*
RACE OF RESPONDENT*
GENDER OF RESPONDENT
交叉制表

文件(F)

编辑(E)

视图(V)

插入(I)

透视(P)

格式(O)

帮助(H)

R USE COMPUTER* RACE OF RESPONDENT* GENDER OF RESPONDENT 交叉制表

GENDER OF RESPONDENT				RACE OF RESPONDENT			
				WHITE	BLACK	OTHER	合计
Female	R USE COMPUTER	YES	计数	336	58	26	420
			RACE OF RESPONDENT 中的 %	77.2%	68.2%	86.7%	76.4%
	NO	计数	99	27	4	130	
			RACE OF RESPONDENT 中的 %	22.8%	31.8%	13.3%	23.6%
	合计	计数	435	85	30	550	
			RACE OF RESPONDENT 中的 %	100.0%	100.0%	100.0%	100.0%
Male	R USE COMPUTER	YES	计数	252	28	23	303
			RACE OF RESPONDENT 中的 %	74.6%	46.7%	76.7%	70.8%
	NO	计数	86	32	7	125	
			RACE OF RESPONDENT 中的 %	25.4%	53.3%	23.3%	29.2%
	合计	计数	338	60	30	428	
			RACE OF RESPONDENT 中的 %	100.0%	100.0%	100.0%	100.0%

图 10-12 枢轴表编辑器

在图 10-12 的菜单栏中, 选择【透视】→【透视托盘 (P)】, 将出现图 10-12 所示的透视托盘。通过该托盘, 可以定制枢轴表的维度。可以交换行和列, 或者把列放到行的位置, 或者把某个变量放到层维度, 反之依然。如图 10-13 所示。

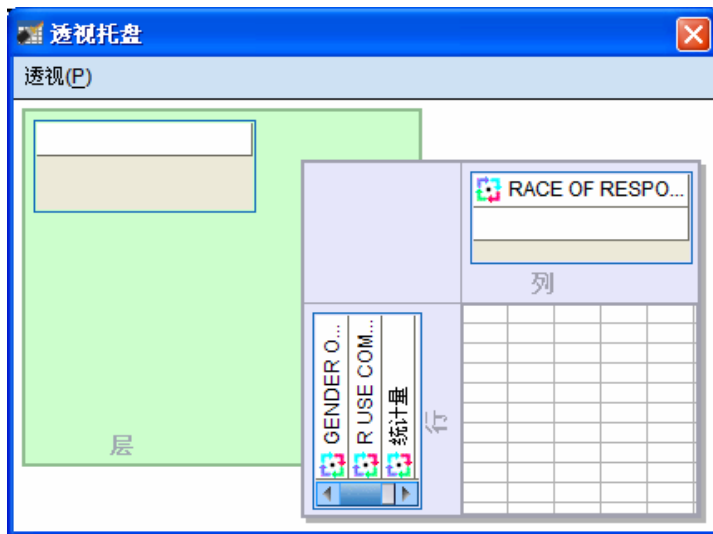


图 10-13 透视托盘

这里, 选择把“GENDER OF RESPONDENT”放到层的维度。在透视托盘中, 用鼠标选中“GENDER OF RESPONDENT”, 然后拖拽到层维度上的空格处, 枢轴表的内容会立即反映透视托盘中维度的变化, 如图 10-14 所示。

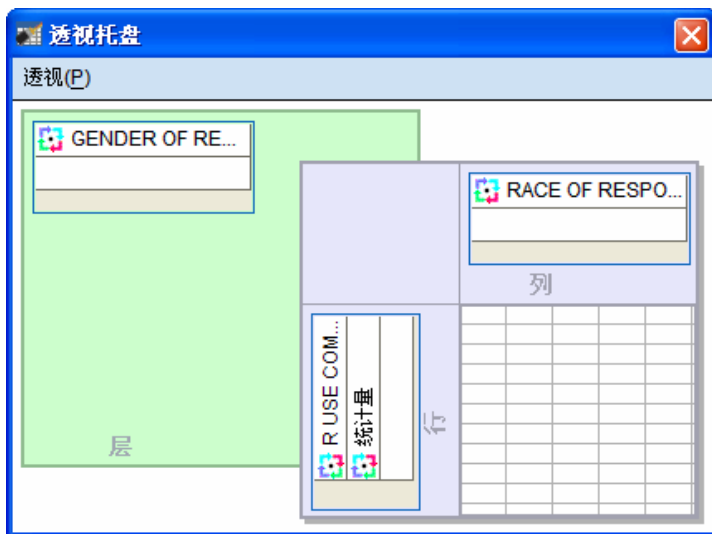


图 10-14 设定维度

这时, 枢轴表编辑器中的表格如图 10-15 所示。在“GENDER OF RESPONDENT”后为一个下拉列表。它列出了该变量的所有取值: 即 Female 和 Male。如果选择 Female, 则列出 Female 组下 RACE OF RESPONDENT 和 R USE COMPUTER 的列联表。如果选择 Male, 则列出 Male 组下 RACE OF RESPONDENT 和 R USE COMPUTER 的列联表。

枢轴表 R USE COMPUTER* RACE OF RESPONDENT* GENDER OF RESPONDENT 交叉制表

文件(F) 编辑(E) 视图(V) 插入(I) 透视(P) 格式(O) 帮助(H)

R USE COMPUTER* RACE OF RESPONDENT* GENDER OF RESPONDENT 交叉制表

GENDER OF RESPONDENT Female

			RACE OF RESPONDENT			
			WHITE	BLACK	OTHER	合计
R USE COMPUTER	YES	计数	336	58	26	420
		RACE OF RESPONDENT 中的 %	77.2%	68.2%	86.7%	76.4%
	NO	计数	99	27	4	130
		RACE OF RESPONDENT 中的 %	22.8%	31.8%	13.3%	23.6%
合计			435	85	30	550
			100.0%	100.0%	100.0%	100.0%

图 10-15 枢轴表

当表格在枢轴表编辑器中单独打开, 或者表格在结果浏览器中其周围有虚线包围, 表格为可编辑状态, 此时可以编辑表格的内容, 例如数值、标签、行标题、列标题等; 也可以设置表格的样式, 例如边框、单元格的属性等。还可以设置表格的外观。

1. 单元格属性

选择【格式】→【单元格属性】，在单元格属性对话框中可以对表格中的每个单元格进行客户化设置，例如单元格中数值的格式、对齐方式、字体等，如图 10-16 所示。



图 10-16 单元格属性设置

2. 表格属性

选择【格式】→【表格属性】，在表格属性对话框中可以对表格的行标签、列标签、标题、边框等进行客户化设置，如图 10-17 所示。



图 10-17 表格属性

3. 表格外观

选择【格式】→【表格外观】，在表格外观对话框中可以选择用户喜欢的外观，如图 10-18 所示。SPSS 的表格外观对话框提供了许多表格模板，它们可以应用到所有的枢轴表。用户也可以选择按钮【编辑外观】，从而创建自己的表格模板。

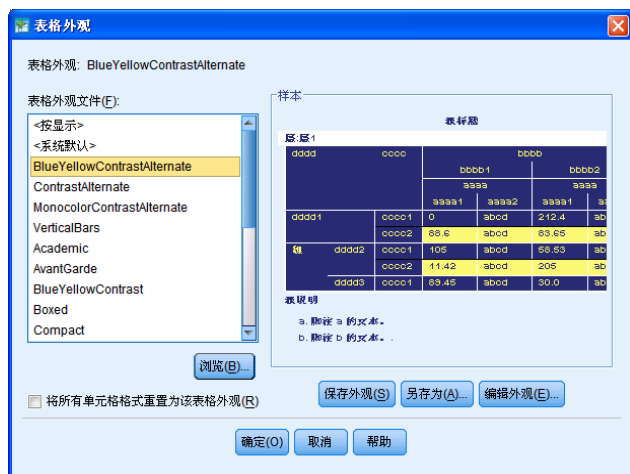


图 10-18 表格外观

10.4 把表格转换为图形

在枢轴表中，可以把枢轴表中的数据“画”出来。在 Ch 10 Sample.spv 中，选择“MARITAL STATUS”表格，然后双击。选择“有效百分比”列的前五个数据，然后单击右键，如图 10-19 所示。这里选择【创建图形】→【饼(E)】，得到如图 10-20 所示的饼图。

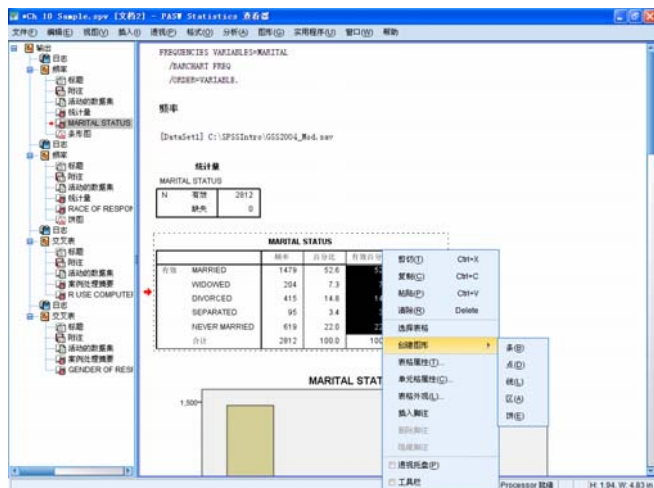


图 10-19 统计表到统计图

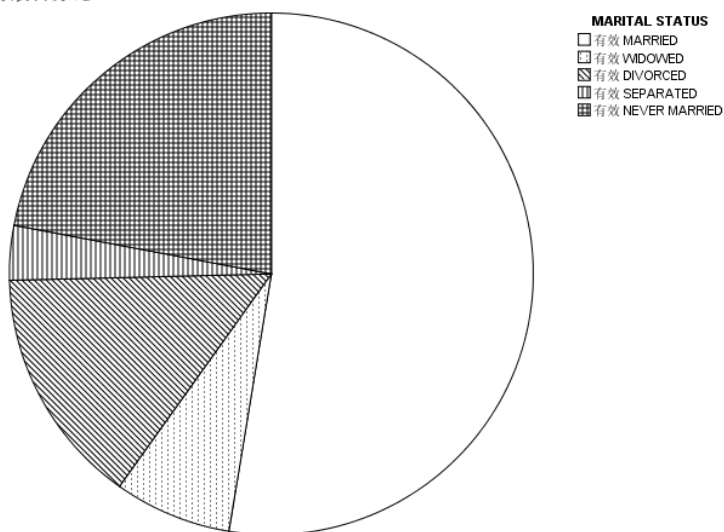
MARITAL STATUS
有效百分比

图 10-20 饼图

10.5 打印输出结果

SPSS 输出管理器中的任何内容都可以直接打印。在正式打印之前，通过【页面设置 (U)】菜单可以调整页面大小、页面方向、添加页眉和页脚等。

选择【文件】→【页面设置】，得到如图 10-21 所示对话框。通过该对话框，用户可以选择纸张的大小、来源、打印的方向和页边距等。

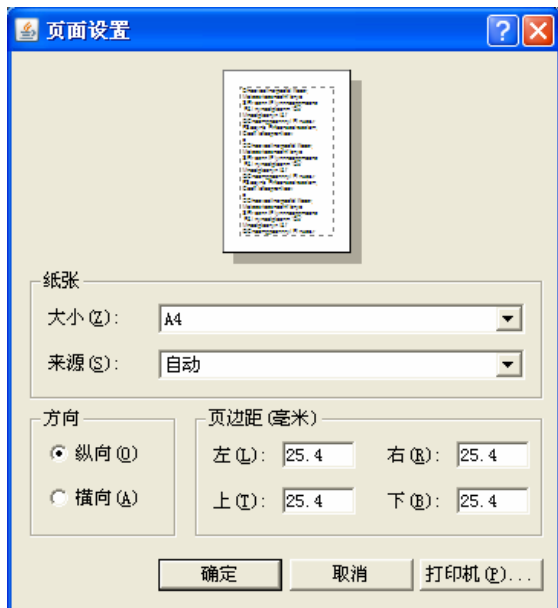


图 10-21 页面设置

选择【文件(F)】→【页面属性】菜单，可以设置打印输出的页眉和页脚，也可以设置打印图表的尺寸、各项之间的空格和页码的起始页等。如图 10-22 所示。



图 10-22 页面属性

10.6 导出输出结果到其他程序

许多不同的应用程序中都可以直接应用 SPSS 的输出结果。例如，可以把 SPSS 输出的图表或者图形包含在报告中。如果输出中有少量的图形或者表格，可以应用复制和粘贴方式直接把 SPSS 输出结果应用到其他应用程序中。

对于较大的输出文件，SPSS 可以把输出结果导出为常见的 MS Word、PDF、HTML、PPT、MS Excel 和文本格式的文件。

10.6.1 直接复制

在 SPSS 结果浏览器中，选择需要复制的内容，然后选择菜单【编辑】→【复制】。转到需要复制到文档，可以是 MS Word、PPT，然后选择菜单【编辑】→【粘贴】，就把 SPSS 输出结果直接复制到了相应的应用程序文档中。例如，可以把结果复制到 MS Word 中，然后选择【自动调整(A)】→【根据窗口调整表格】或者【自动调整(A)】→【根据内容调整表格】，则表格的内容将自动调整以最

佳方式适应相应的文档。

10.6.2 导出为其他文件格式

当导出 SPSS 输出结果为 Excel 或者文本格式时，表格不能够被导出。选择菜单【文件】→【导出】菜单，得到如图 10-23 所示的导出结果。



图 10-23 导出结果

“导出的对象”：设定导出的内容。

- 全部（A）：导出所有输出结果，不论是显示状态还是隐藏状态。
- 所有可见（V）：只导出显示状态的输出结果，忽略隐藏状态的结果。
- 选定（D）：只导出选定的输出结果

“类型（T）”：设定导出文件的类型，有 7 中不同的文件类型，包括 MS Word 格式（*.doc）、*.HTML、*.pdf、*.PPT、文本格式文件（包括纯文本、UTF8 和 UTF16 三种格式）。

“更改选项（C）”：设置枢轴表的输出方式。

10.7 小结

本章主要介绍 SPSS 结果浏览器的结构、应用等。它们包括浏览和编辑 SPSS 的分析结果、移动、复制和删除结果、编辑文本等。同时，本章介绍了枢轴表编辑器、学习了如何把表格转换为图形，最后介绍了如何把 SPSS 输出结果导出为其他格式的应用程序。

SPSS 编程简介

本章学习目标:

- 了解应用 SPSS 的几个不同阶段;
- 了解 SPSS 语法功能及获取方式;
- 了解 SPSS 语法编辑器;
- 了解 SPSS 语法命令的特点。

SPSS 作为一个强大的数据分析工具,提供了许多把复杂的数据分析过程自动化的工具和手段。SPSS 的脚本编程工具结合了 Microsoft 的 Visual Basic 和 COM 模型,使用户的数据分析可以实现自动化。用户进行相同的分析时,不要再重复第一次的复杂过程;另外,复杂的分析过程可以通过简单的点击即可完成。分析人员可以结合他们的专家经验和知识,把他们的分析过程写成脚本。这样,更多没有经验的用户就可以应用他们的脚本,通过简单的点击来完成复杂的分析过程。

- 脚本编程工具极大的扩展了 SPSS 的分析过程,给分析带来极大的灵活性;
- 更多的自动化。比如,许多复杂的 SPSS 分析过程可能需要周期性的执行,脚本极大的减轻了重复编写分析过程的复杂过程。
- 更多的集成。可以和其他的应用程序,例如 Word, Excel, Powerpoint, PDF 等集成。
- 更广泛的应用。没有 SPSS 知识的用户也可以应用 SPSS 的分析过程。
- 可以应用更广泛的分析过程。用户可以编写自己的 SPSS 中没有的分析过程,或者修改 SPSS 的分析过程,或者直接借用其他的分析过程,然后把他们集成到 SPSS 的分析环境中。

获得以上的方便性和灵活性的代价是,必须知道如何在 SPSS 中编程。最早的 SPSS 编程语言是 SPSS 的命令语法 (Command Syntax)。除此之外, SPSS 还提供了脚本编程语言,一种脚本编程语言为 Basic 语言。SPSS 中的 Basic 语言叫做 Sax Basic,它和 Microsoft 的 Visual Basic 是兼容的。因此把 SPSS 的 Basic 脚本和 Microsoft

的 VBA 结合是十分容易的。

11.1 应用 SPSS 的五阶段

应用 SPSS 可分为以下五个阶段进行。

第一阶段 通过菜单和对话框，交互式应用 SPSS

大多数 SPSS 用户，甚至大部分的 SPSS 专家也是从菜单和对话框开始一项新的分析任务。通常，我们通过鼠标单击，打开数据文件，选择要分析的变量，通过菜单设置分析选项。中间，我们偶尔要输入数值，例如在 K-means 聚类中，要手工输入聚类数量。用图形菜单方式进行频率分析，需要选择要分析的变量、设定需要输出的统计量，如图 11-1 所示。

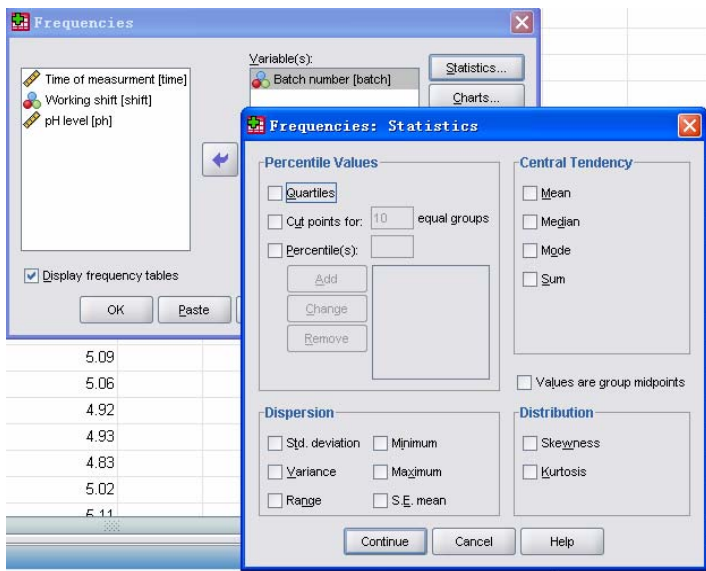


图 11-1 通过对话框来定义——项频数分析

第二阶段 通过命令语法，交互式应用 SPSS

如果要经常进行某些分析过程，可能在不同的数据集合上应用相同的分析。这时候，可以把分析过程用 SPSS 的语法命令记录下来。简单地把相应于变量和分析选项的语法记录下来，然后粘贴到语法文件中（.SPS 后缀），或者包含其他语法命令，生成 SPSS 作业。SPSS 语法命令可以整体执行，也可以执行其中的某一部分，如图 11-2 所示。

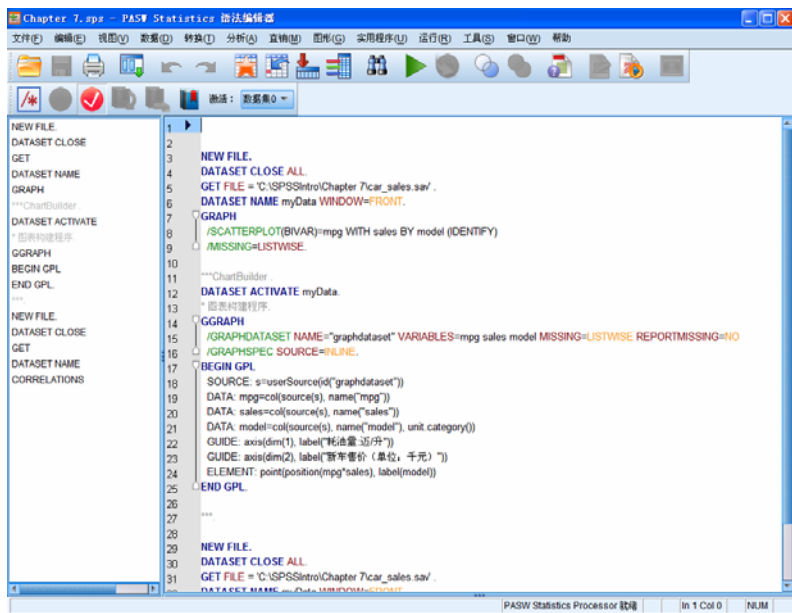


图 11-2 SPSS 18 命令语法编辑器

第三阶段 通过命令语法，应用 SPSS 批量作业模式

可以通过 SPSS 的批量作业模式，该模式运行 SPSS 语法命令文件，输出一个命名的 SPSS 输出文件（.SPO）。该模式基本没有用户和 SPSS 系统的交互过程，如图 11-3 所示。



图 11-3 SPSS 生产工作

第四阶段 SPSS 内部脚本

从第三阶段的 Syntax 可以看出 Syntax 让用户高效的完成用菜单完成的所有分析工作（尤其是对输入而言）。但是 SPSS 脚本可以完成许多 Syntax 不能完成的工作。例如，

可以用脚本操作和控制输出。比如，脚本可以检查输出的数值。基于输出值，采取其他的步骤，例如如果前边的正态测试显著，可以运行 t 检验。

脚本可以容易得到外部信息，并用它们来执行分析。比如，它可以找出用户的临时文件目录，并用它来存放自动化过程中产生的临时文件，分析任务完成后删除用不到的部分。

脚本可以和其他的应用通信，比如，它可以产生一个输出，然后自动的复制到 Word 或者 PowerPoint 中。

第五阶段 SPSS 外部脚本

第五阶段在拥有上一个阶段的所有优点的基础上，给程序更多的自动化，比如用户开发的界面或者应用。例如，下面用户自己开发的对话框，要求输入时间范围和符合条件的记录数目，然后在输出中显示报告。SPSS 自定义对话框，如图 11-4 所示。



图 11-4 SPSS 自定义对话框

该应用对一个 Excel 数据文件中的一系列基于时间点的数据，在 SPSS 中进行分析，然后把输出导出到 Excel 表格中。尽管实际的分析过程十分的复杂，但是我们的应用程序封装了中间过程，我们的应用对最终用户是透明的，简单易用。

图 11-5 是一个 SPSS 脚本程序的示例界面，该脚本重新设置 SPSS 内部随机数的种子为 1234（默认为 20000），然后在 SPSS 的输出窗口中检查系统的随机数种子和其他的设置。

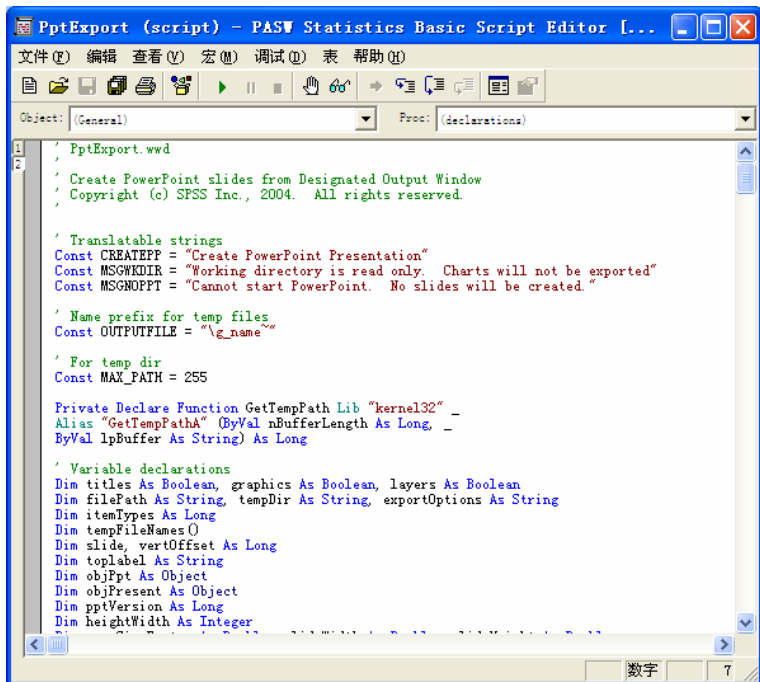


图 11-5 SPSS 脚本程序

11.2 SPSS 语法简介

SPSS 语法是由一系列命令组成的 SPSS 指令，它指导 SPSS Statistics 如何管理、修改和分析数据。SPSS 的语法命令自 SPSS 软件伊始就有了，二十多年之后，才有了图形用户界面。当 SPSS 在 20 世纪 70 年代作为主流的统计分析软件，必须学习语法命令才能应用。那时候，分析者需要熟悉语法命令的通用规则、语法、关键词。到 90 年代初，SPSS 开发了图形用户界面，把用对话框来生成语法命令程序，而用户则不必直接和语法命令程序打交道。

当前用户应用 SPSS 的方式主要有两种，通过具有菜单和对话框的图形用户界面应用 SPSS 或者通过语法命令来应用 SPSS。现在大部分的 SPSS 用户都直接应用图形用户界面；甚至好多用户都不知道 SPSS 可以编程。对于高校里学习统计学或者相关的课程，由于课时量的限制，教师甚少深入地去讲 SPSS 编程。对于一次应用，以后不再接触的用户，图形用户界面是最好的选择，他们没有必要去学习另外一种效率高的应用方式。许多专业人士仍然坚持通过 SPSS 的命令语法来应用 SPSS 进行数据分析。

SPSS 语法命令具有下列图形用户界面的菜单方式所不具有的优点：

- 重复性的任务；
- 更多的功能选择；
- 有些分析方式只能通过语法命令实现；
- 分析过程存档；
- 方便和其他人进行交流。

11.3 SPSS 语法编辑器

11.3.1 图形用户界面和语法编程相结合

SPSS 的命令语法具有很好的存档分析过程的功能。对于一些需要重复进行的分析过程，建议以语法命令的方式记录下分析的过程，这对于不需要应用 SPSS 命令语法的用户有时也是很有帮助的。

SPSS 可以把图形用户界面的菜单和对话框过程记录下来。例如，对于第一章用到的 `Employ Data.sav`，如果需要对雇佣类别进行频率分析，可以通过选择菜单【分析】→【描述统计】→【频率 (F)】，SPSS 图形用户界面如图 11-6 所示。得到如图 11-7 所示的频率对话框后，在该对话框中设置相应的选项，然后单击【确定】按钮即可。



图 11-6 SPSS 图形用户界面



图 11-7 SPSS 对话框设置

SPSS 可以完全记录下该操作过程。在完成图 11-7 所示的设置后,不要单击【确定】按钮,而是单击【粘贴(P)】按钮,则 SPSS 把上述操作过程以语法命令形式写入到语法编辑器中,如图 11-8 所示。

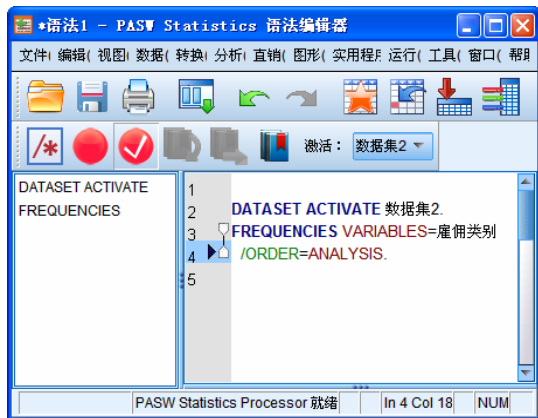


图 11-8 SPSS 语法编辑器

11.3.2 SPSS 操作日志

另外一种记录 SPSS 操作过程的方法是保留 SPSS 的日志文件。SPSS 可以记录下所有的操作过程,如图 11-9 所示。双击输出查看器中左端的“日志”,右边将显示其下面所输出结果的命令语法程序。同时,该部分也会显示程序运行中的警告信息或者出错信息。

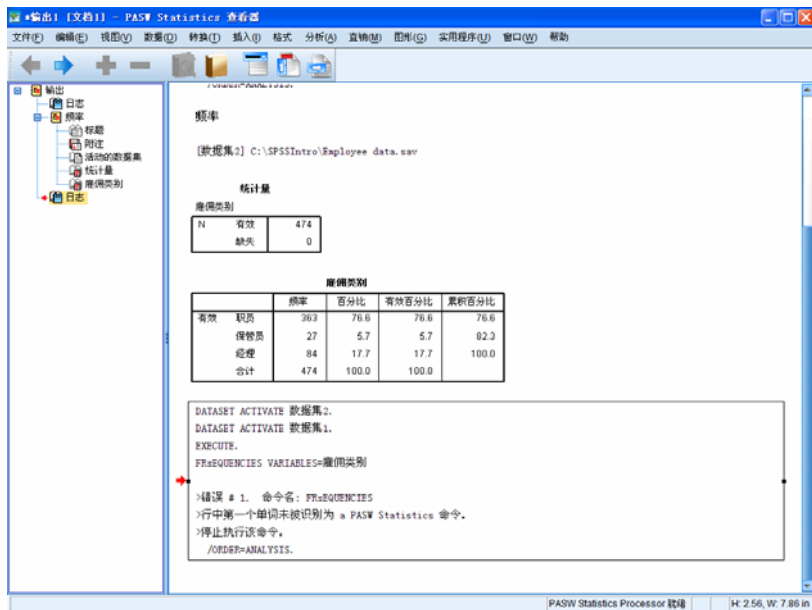


图 11-9 SPSS 日志文件

在 SPSS 的选项部分可以设置是否保留操作过程的日志。选择【编辑(E)】→【选项(N)】，进入“文件位置”标签项。如图 11-10 所示。在“会话日志”部分可以设置是否保留日志文件以及日志文件的位置。

- 附加(P)：新的操作日志将添加在以前的操作日志文件的尾部。
- 覆盖(T)：新的操作日志将覆盖以前的操作日志文件。



图 11-10 设置日志文件的保存方式

11.3.3 SPSS 语法编辑器简介

SPSS 语法编辑器是一个功能较强的语法编辑集成环境，它既可以编辑语法命令，也可以运行语法命令程序。除了具有【工具(S)】菜单和它相应的工具按钮以外，该编辑器和数据视图的菜单和工具按钮大部分相同。工具菜单提供了语法编辑器的设置、语法程序断点的设置和清除、设置和切换书签等子菜单，这些子菜单的一部分可以在工具按钮中找到，如图 11-11 所示。

1. 编辑器窗口

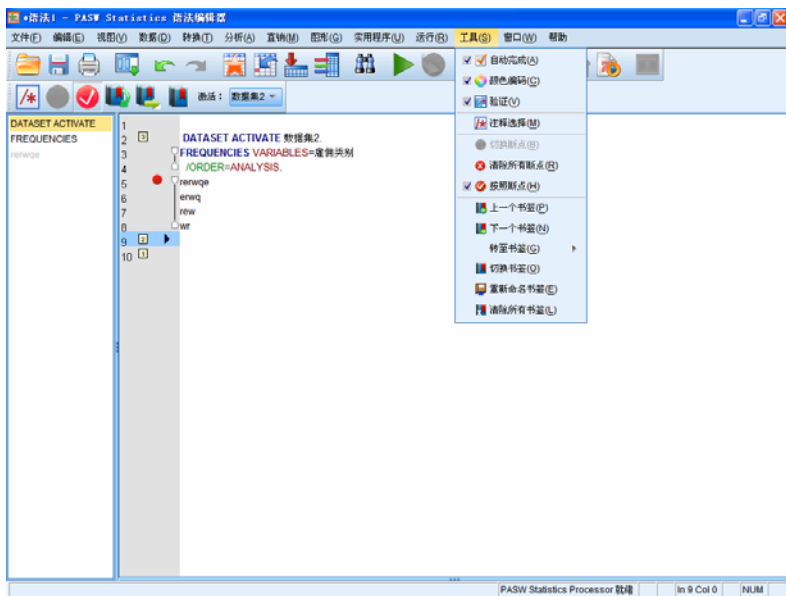


图 11-11 SPSS 语法命令编辑器

SPSS 语法编辑器窗口可以分为三个区域：导航区域、语法编辑区域和调试信息区域，如图 11-12 所示。

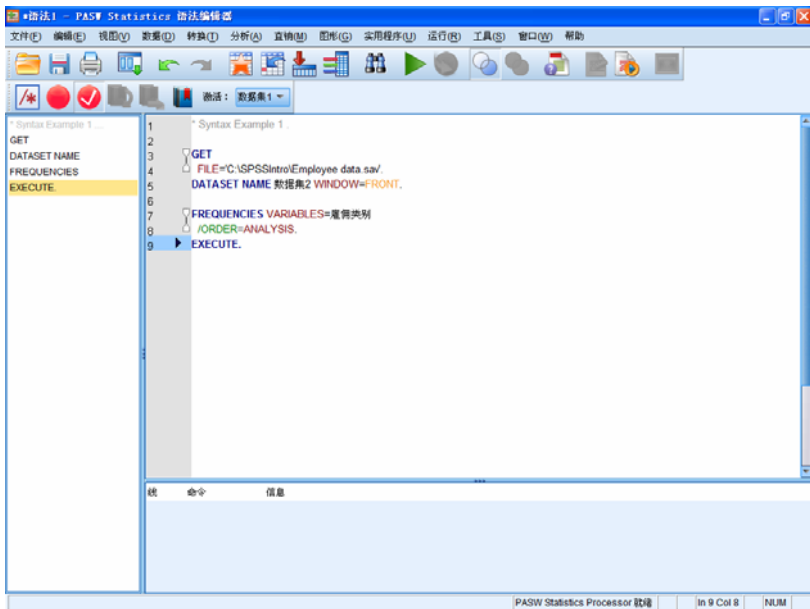


图 11-12 SPSS 语法编辑器窗口

- 导航区域：图 11-12 的最右端的为导航窗口，它类似于一本书的目录，列出了语法编辑区的语法命令的所有语句开始的关键词，通过双击关键词可以快速地定位到右侧语法文件中的相应部分。它是定位大型的语法程序中相应语句的高效工具。

- 语法编辑区域：图 11-12 的上部分为语法编辑区域，在该部分可以输入、编辑语法命令。该区域的最左端呈灰色的区域用于显示程序的行号、书签、断点和命令组等信息。在该区域可以通过单击鼠标右键来选择显示或隐藏相应的内容。
- 调试信息区域：右下角命令调试信息，如果语法命令有错误，运行时将在该区域显示出错的命令和相应的调试信息，用户可以基于此来调试程序，直到正确运行。

2. 编辑器应用

如图 11-13 所示，当在语法命令编辑器中输入命令的一部分时，SPSS 会智能的选出可能的命令供用户选择，这样一方面节省了时间，另外避免了由于用户拼写命令错误而导致的语法错误。该功能即为高级程序集成环境中的“自动完成”功能。

在编辑器中输入方差分析关键字 ANOVA 时，当输入到第三个字母“ANO”时，右下方将弹出一个包含输入字母的相关命令以供用户选择。

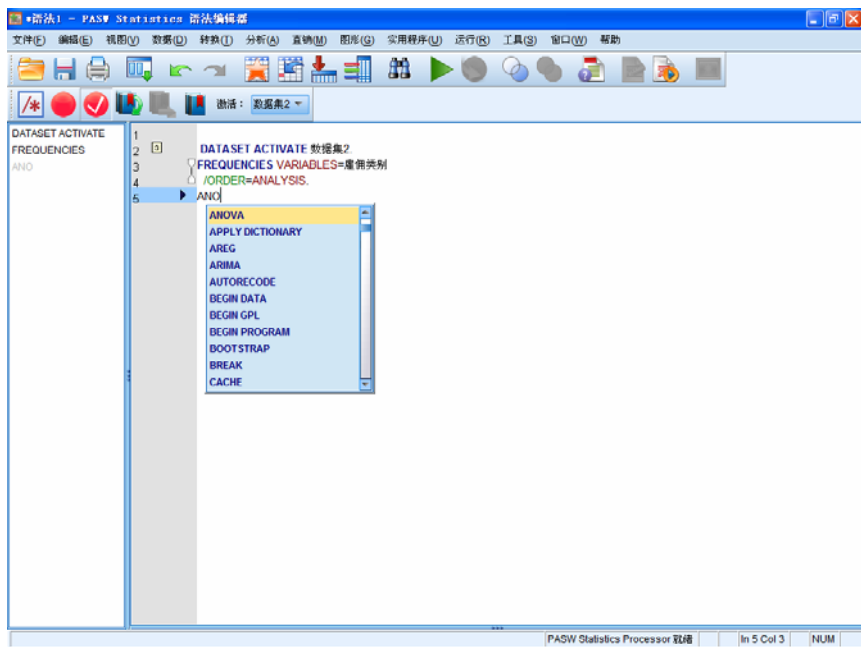


图 11-13 SPSS 命令语法的自动完成

默认情况下，正确的命令为蓝色，子命令为绿色，关键词（即子命令的取值设置）为褐色，变量名为黑色。用户可以根据自己的爱好来改变以上设置。可以通过选择【编辑（E）】→【选项（N）】，进入“语法编辑器”标签项来修改相应的设置，如图 11-14 所示。



图 11-14 命令语法程序文字外观显示设置

11.4 应用 SPSS 语法命令进行编程

SPSS 所有的语法命令都是以一个命令开始的，一个命令单位结束之后，需要有一个结束符合来“.”。一般情况下，一个命令下面有多个子命令来对开始的命令进行设置。可以这样理解，最开始的命令是方法性的，它决定分析方法的思路类型；而子命令则是对具体命令的喜欢和完善。基本要求如下：

- 每个命令必须从新一行的第一列开始，并以点号结束 (.) (小数点)；
- 大多数的子命令以斜线 (/) 分隔；
- 必须完整地拼写变量名；
- 省略号或引号中包括的文本必须位于一行中；
- 续行必须至少有一个空格的缩进；
- 每个命令结尾处的点号是可选的。

例如，要分析 Employ data 中变量 Jobcat 和 Gender 的频率，可以输入下列程序。

```
FREQUENCIES
  VARIABLES=JOB CAT  GENDER
  /PERCENTILES=25 50 75
  /BARCHART.
```

这里 FREQUENCIES 是主要命令，表示进行频率分析。其后的 VARIABLES 为关键字，它用于指定待分析的变量。带斜杠的命令 PERCENTILES 和 BARCHART

为子命令，表示在对变量 **JOB**CAT、**GENDER** 进行频率分析的同时显示它们的 25、50 和 75 三个百分位数，另外做出它们的条形图。

11.5 小结

本章简单介绍了 **SPSS** 的各种编程方式，它包括命令语法、脚本程序等。重点介绍了 **SPSS** 命令语法的特点，以及 **SPSS** 语法命令编辑器。语法命令可以帮助用户养成保留分析过程的习惯。这对于以后的数据分析，和同事或者同行交流分析结果提供了便利。**SPSS** 的语法编辑器正在向商业的高级程序集成环境发展，用户需要熟悉该编辑器的功能与特点。

思考与练习

1. 以下哪些是 **SPSS** 语法命令所不能够完成的：
 - A) **SPSS** 语法命令可以自动对 **SPSS** 生成的结果进行分析
 - B) **SPSS** 语法命令可以大大简化重复性的任务
 - C) **SPSS** 语法命令可以提供菜单窗口所不能够提供的功能选择
 - D) **SPSS** 语法命令可以提供分析过程的文档以方便与其他人进行交流
2. 有关 **SPSS** 语法命令的论断，错误的是：
 - A) 每个命令必须从新的一行开始并且必须以点号结束 (.)
 - B) 可以采用简化的变量名，例如变量名的前三个字母
 - C) 子命令之间必须用分号分隔 (;)
 - D) 每个命令结尾处的点号是可选的
 - E) 每一行语法命令的长度不能超过 256Byte.
3. 以下哪些功能是 **SPSS 18** 的语法命令编辑器所不具备的：
 - A) 可以用颜色来区分不同的语法部分
 - B) 上下文敏感功能可以自动完成输入的关键字
 - C) 可以自动生成语法命令的框架
 - D) **SPSS** 的菜单窗口的操作过程可以被录制在语法编辑器中
4. 哪些可能是语法命令 “**COMPUTE results=Units<2.**” 的结果：
 - A) 生成一个新的变量 **results**。如果 **Units<2**，则 **results** 的值和 **Units** 的值一样，否则为缺失值
 - B) 生成一个新的变量 **results**，它的值和 **Units** 的值一样
 - C) 生成一个新的变量 **results**。如果 **Units<2**，则 **results** 的值为 1，否则为 0
 - D) 以上都不正确

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396；(010) 88258888

传 真：(010) 88254397

E-mail: dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036



《SPSS 18 数据分析基础与实践》读者交流区

尊敬的读者：

感谢您选择我们出版的图书，您的支持与信任是我们持续上升的动力。为了使您能通过本书更透彻地了解相关领域，更深入的学习相关技术，我们将特别为您提供一系列后续的服务，包括：

1. 提供本书的修订和升级内容、相关配套资料；
2. 本书作者的见面会信息或网络视频的沟通活动；
3. 相关领域的培训优惠等。

请您抽出宝贵的时间将您的个人信息和需求反馈给我们，以便我们及时与您取得联系。

您可以任意选择以下三种方式与我们联系，我们都将记录和保存您的信息，并给您提供不定期的信息反馈。

1. 短信

您只需编写如下短信：B11255+您的需求+您的建议

发送到1066 6666 789（本服务免费，短信资费按照相应电信运营商正常标准收取，无其他信息收费）
为保证我们对您的服务质量，如果您在发送短信24小时后，尚未收到我们的回复信息，请直接拨打电话（010）88254369。

2. 电子邮件

您可以发邮件至jsj@phei.com.cn或editor@broadview.com.cn。

3. 信件

您可以写信至如下地址：北京万寿路173信箱博文视点，邮编：100036。

如果您选择第2种或第3种方式，您还可以告诉我们更多有关您个人的情况，及您对本书的意见、评论等，内容可以包括：

- （1）您的姓名、职业、您关注的领域、您的电话、E-mail地址或通信地址；
- （2）您了解新书信息的途径、影响您购买图书的因素；
- （3）您对本书的意见、您读过的同领域的图书、您还希望增加的图书、您希望参加的培训等。

如果您在后期想退出读者俱乐部，停止接收后续资讯，只需发送“B11255+退订”至10666666789即可，或者编写邮件“B11255+退订+手机号码+需退订的邮箱地址”发送至邮箱：market@broadview.com.cn亦可取消该项服务。

同时，我们非常欢迎您为本书撰写书评，将您的切身感受变成文字与广大书友共享。我们将挑选特别优秀的作品转载在我们的网站（www.broadview.com.cn）上，或推荐至CSDN.NET等专业网站上发表，被发表的书评的作者将获得价值50元的博文视点图书奖励。

我们期待您的消息！

博文视点愿与所有爱书的人一起，共同学习，共同进步！

通信地址：北京万寿路 173 信箱 博文视点（100036） 电话：010-51260888

E-mail: jsj@phei.com.cn, editor@broadview.com.cn

www.phei.com.cn
www.broadview.com.cn